QARAB: A Question Answering System to Support the Arabic Language

Bassam Hammo Hani Abu-Salem Steven Lytinen

DePaul University
School of Computer Science, Telecommunications and Information Systems
243 S. Wabash Avenue, Chicago IL 60604

bhammo@condor.depaul.edu habusalem@cti.depaul.edu lytinen@cs.depaul.edu

Martha Evens
Illinois Institute of Technology
Computer Science Department
10 West 31st Street, Chicago, IL 60616
evens@iit.edu

Abstract

We describe the design and implementation of a question answering (QA) system called QARAB. It is a system that takes natural language questions expressed in the Arabic language and attempts to provide short answers. The system's primary source of knowledge is a collection of Arabic newspaper text extracted from Al-Raya, a newspaper published in Qatar. During the last few years the information retrieval community has attacked this problem for English using standard IR techniques with only mediocre success. We are tackling this problem for Arabic using traditional Information Retrieval (IR) techniques coupled with a sophisticated Natural Language Processing (NLP) approach. To identify the answer, we adopt a keyword matching strategy along with matching simple structures extracted from both the question and the candidate documents selected by the IR system. To achieve this goal, we use an existing tagger to identify proper names and other crucial lexical items and build lexical entries for them on the fly. We also carry out an analysis of Arabic question forms and attempt a better understanding of what kinds of answers users find satisfactory. The paucity of studies of real users has limited results in earlier research.

1 Introduction

In recent years, there has been a marked increase in the amount of data available on the Internet. Users often have specific questions in mind, for which they hope to get answers. They would like the answers to be short and precise, and they always prefer to express the questions in their native language without being restricted to a specific query language, query formation rules, or even a specific knowledge domain. The new approach taken to matching the user needs is to carry out actual analysis of the question from a linguistic point of view and to attempt to understand what the user really means.

QARAB is the result of coupling traditional Information Retrieval (IR) techniques with a sophisticated Natural Language Processing (NLP) approach. The approach can be summarized as follows: the IR system treats the question as a query in an attempt to identify the candidate

documents that may contain the answer; then the NLP techniques are used to parse the question and analyze the top ranked documents returned by the IR system.

Natural Language Processing (NLP) in the Arabic language is still in its initial stage compared to the work in the English language, which has already benefited from the extensive research in this field. There are some aspects that slow down progress in Arabic Natural Language Processing (NLP) compared to the accomplishments in English and other European languages [Al-Daimi & Abdel-Amir, 1994]. These aspects include:

- Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task.
- The absence of diacritics (which represent most vowels) in the written text creates ambiguity and therefore, complex morphological rules are required to identify the tokens and parse the text.
- The writing direction is from right-to-left and some of the characters change their shapes based on their location in the word.
- Capitalization is not used in Arabic, which makes it hard to identify proper names, acronyms, and abbreviations.

In addition to the above linguistic issues, there is also a lack of Arabic corpora, lexicons, and machine-readable dictionaries, which are essential to advance research in different areas.

2 Background

Advances in natural language processing (NLP), information retrieval techniques (IR), information extraction (IE), as well as the computer industry, have given QA a strong boost. Modern question-answering systems have started incorporating NLP techniques to parse natural language documents, extract entities and relations between entities, resolve anaphora, and other language ambiguities [Harabagiu et al., 2000; Vicedo & Ferrández, 2000].

Research in Question-Answering (QA) is not new. The QA problem has been addressed in the literature since the beginning of computing machines. The AI/NLP communities initiated traditional work to address question-answering using structural methods. Early experiments in this direction implemented systems that operate in very restricted domains (e.g. SHRDLU [Winogard, 1972] and LUNAR [Woods, 1972]). In the QUALM system, Lehnert [1978] took a further step, based on the conceptual theories of Schank & Abelson [1977], to understand the nature of the questions and classify them in a way similar to how human beings understand and answer questions. SCISOR [Jacobs & Rau 1990] aimed at question answering and text extraction more than information retrieval. It combined natural language processing, knowledge representation, and information retrieval techniques with lexical analysis and word-based text searches. The MURAX system [Kupiec, 1993] used robust linguistic methods to answer closed-class natural language questions. It presented the user with relevant text in which noun phrases are marked. A less automated approach like Ask Jeeves [1996] approached the QA problem by pointing the questioner to Web links that might contain information relevant to the answer to the question. Ask Jeeves benefited from advanced natural language processing techniques combined with data mining processing and a huge expanding knowledge base. Another system, with a different approach, is the FAQFinder system [Burke et al., 1997], which attempted to solve the question-answering problem using a database of question-answer pairs built from existing frequently asked question (FAQ) files. Two other important systems are the START system [Katz, 1997], which is based on annotations from the Web and the Q&A system [Budzik & Hammond, 1999], which is a semiautomatic, natural language question-answering and referral system. The system is based on a huge knowledge base and human experts who volunteered their time to respond to the users' questions.

Recently, attention has begun to be focused on developing question-answering systems that do

not rely on a knowledge base and that can fetch answers from huge unstructured text. New QA systems enhanced with NLP and IR techniques have been developed to extract textual answers for open-domain questions and provide a framework for modern information retrieval [TREC-8, 1999; TREC-9, 2000].

The overall aim of this QA track was to retrieve small pieces of text that contain the actual answer to the question rather than the list of documents traditionally returned by retrieval engines [Voorhees & Tice, 2000]. The TREC-8 QA track attracted researchers from both industry and academia. Twenty organizations participated in this track with different approaches and their systems were evaluated. The participating systems were tested on a huge set of unstructured documents and a set of fact-based questions.

Generally speaking, most of the TREC-8 long-string answer (250-bytes) participants attempted to solve the QA problem from the information retrieval (IR) point of view by locating the most relevant documents from the collection and then extracting the sentences most relevant to the query from the documents just located. The systems relying on this *"bag-of-words"* approach (e.g. [Allan et al., 1999]; [Cormack et al., 1999]; [Lin & Chen, 1999]; [Shin et al., 1999] and the passage-retrieval run of AT&T [Singhal et al., 1999]) deal with the question without considering its grammatical or semantic characteristics and they apply conventional IR techniques to extract the answer. Even though the *"bag-of-words"* approach was commonly used in TREC-8, the systems based on this approach were inadequate to handle the short-string (50-byte) answers.

On the contrary, the short string (50-byte) participants (e.g. [Breck et al., 1999]; [Ferret et al., 1999]; [Hull, 1999]; [Humphreys et al., 1999]; [Litkowski, 1999]; [Moldovan et al., 2000]; [Oard et al., 1999]; [Singhal et al., 1999]) agreed on the importance of applying several natural language processing techniques to solve the problem. Among these techniques are: part-of-speech tagging, shallow parsing, query type identification and named entity recognition. Because the number of test documents to be analyzed for each query was huge, the majority of the systems in this band used the "bag-of-words" approach as an initial step to retrieve the relevant passages that contain the possible answer. Another approach to the QA problem combines IR techniques with Information Extraction (IE) techniques for extracting named entities, e.g., [Ogden et al., 1999]; [Takaki, 1999]; and [Srihari & Li, 1999]. A detailed description of the track and the results are available at [Voorhees & Tice, 1999].

It is obvious from the increasing number of systems participating in TREC-9 and the worldwide interest in this research area that Question Answering is the most promising framework for finding answers to natural language questions from a huge amount of textual data. Cardie et al. [2000] pointed out that building *"open-ended question answering systems that allow users to pose questions of any type and in any language, without domain restrictions, is still beyond the scope of any QA system today"* (p. 180). Harabagiu et al. [2000] indicated that advanced tools (such as dialog understanding and text mining) are essential for the success of future QA systems. Until the advanced tools are implemented, she suggested that we keep approximating the complexity of Question Answering with NLP enhancements of IR and IE techniques [Harabagiu et al., 2000].

3 QARAB System

3.1 Overview

In the last decade, the volume of Arabic textual data has started growing on the Web and Arabic software for browsing the Web is improving. Unfortunately, much of the earlier Arabic text available on the Web was posted as images, which makes it unsuitable for search or processing. As of today, there is an increase in the amount of Arabic textual material available on the Web in the form of news articles and books.

The main goal of the QARAB system is to identify text passages that answer a natural language question. The task can be summarized as follows: *Given a set of questions expressed in*

Arabic, find answers to the questions under the following assumptions:

- The answer exists in a collection of Arabic newspaper text extracted from the Al-Raya newspaper published in Qatar.
- The answer does not span through documents (i.e. all supporting information for the answer lies in one document)
- The answer is a short passage.

The basic QA processing in QARAB is composed of three major steps:

- Processing the input question
- Retrieving the candidate documents (paragraphs) containing answers from the IR system
- Processing each one of the candidate documents (paragraphs) in the same way as the question is processed and returning sentences that may contain the answer.

The QARAB system will be evaluated over a wide range of question types provided by Arabic users during the testing and the final phases. The same users will then assess whether the answers produced by the system are satisfactory.

3.2 QARAB Structure

The complete QARAB system is depicted in Figure 1; it has the following overall structure:

3.2.1 The IR System

The IR system, which we are implementing from scratch, is based on Salton's vector space model [Salton, 1971]. First, it processes the text collection from the Al-Raya newspaper and constructs an inverted file system, from which the answers to the natural language questions will be extracted. The purpose of the IR system is to search the document collection to select documents containing information relevant to the user's query.

Implementing the Information Retrieval System

Information Retrieval (IR) systems can be constructed in many various ways. Lundquist et al. [1999] proposed an Information Retrieval (IR) system that can be constructed using a relational database management system (RDBMS). Our IR system is depicted in Figure 2 and it contains the following database relations:

- ROOT_TABLE (*Root_ID, Root*) – to store the available distinct roots of the terms extracted from the Al-Raya document collection (one row per root).
- STEM_TABLE (*Stem_ID, Root_ID, Stem, Document_Frequency, IDF*) – to store all distinct stems from the document collection. The stem frequency in the entire document collection and the inverse document frequency of each stem are calculated and stored (one row per stem).
- POSTING_TABLE (*Posting_ID, Stem_ID, Document_ID, Paragraph_ID, Position, Length*) – to store all the occurrences of the stems extracted from the entire document collection (one row per stem).
- DOCUMENT_TABLE (*Document_ID, Document_Title, Document_Date, Document_Path*) — to store document information (one row per document)
- PARAGRAPH_TABLE (*Paragraph_ID, Document_ID, Paragraph*) — to store all the paragraphs extracted from the document collection (one row per paragraph). This speeds up the analysis and the processing of the relevant passages that might answer to the user's question.
- QUERY_TABLE (*Word, Weight*) – to store query information. This includes the original query words and the set of expanded words. The set of expanded words is obtained by extracting the available roots of the original query words, finding their equivalent Root_ID's in the ROOT_TABLE, and then finding their corresponding terms stored in the STEM_TABLE. The weight of each word is calculated and stored (one row per word).

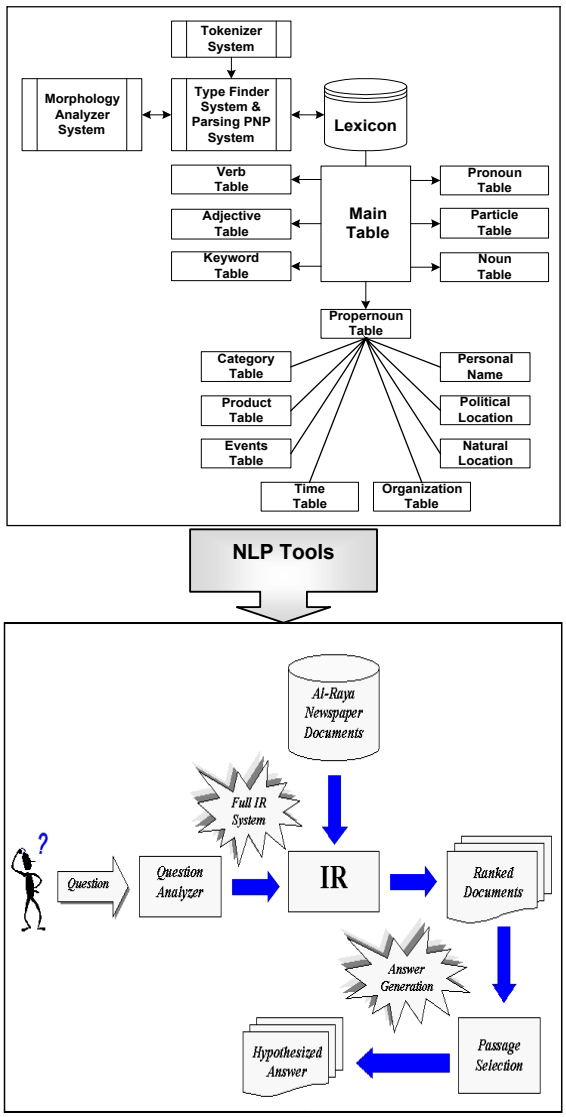

Figure 1. System Components

3.2.2 The NLP System

The second component of the system (the NLP system) shown in Figure 1 was implemented by Abuleil [1999] to experiment in building a large Arabic lexicon. The NLP system is composed of a set of tools to tokenize and tag Arabic text, identify some features of the tokens and, most important, to identify proper names. The following is a description of the overall structure and functionality of the NLP system.

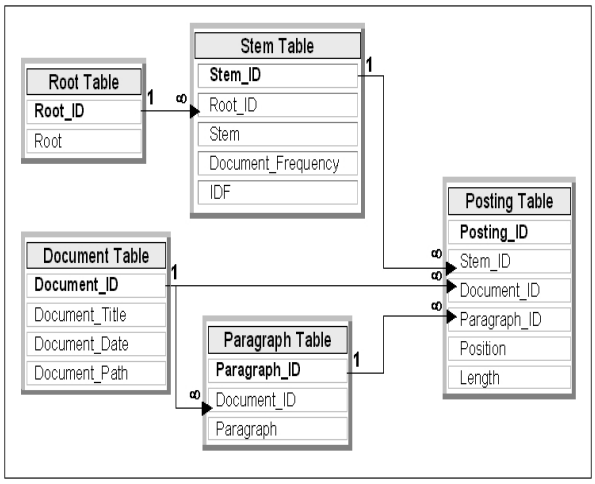

Figure 2. Relational Database Information Retrieval System

The tagger was designed to construct a comprehensive Arabic lexicon. The system is used to parse Arabic words and determine their parts of speech (verbs, nouns, particles). Also it is used to figure out the features of each word (gender, number, person, tense), mark proper nouns in the text and determine their types (personal names, locations, organizations, times, dates, etc.).

The NLP system comprises the following modules:

- The *tokenizer*, which is used to extract the tokens.
- The *type finder*, which is used to assign a part-of-speech to each token.
- The *feature finder*, which is used to determine the features of each word.
- The *proper noun phrase parser*, which is used to mark proper nouns.

The *type finder* module starts a lexicon lookup process for each token. When there is an unknown word in the text, the system can apply the proper noun phrase parser to tag the word as a proper noun. The recognition process occurs in multiple stages in which a list of patterns and heuristics may be applied to mark the proper noun. When the word is tagged as a proper noun, it is added automatically to the lexicon with all its possible

features. Being able to identify the proper names, among other actual entities, in the text is an important step in understanding and using the text. Unfortunately, this is not a straightforward task in Arabic as it is in English and most European languages since the uppercase/lowercase distinction does not exist in Arabic text. Thus, we have to learn more about the common patterns in which these entities occur in Arabic contexts.

4 The Basic Outline of Processing in the IR System

4.1 Document Processing

This step is essential for our system. First, the newspaper articles from the Al-Raya newspaper are saved in text format using the *Arabic Windows 1256 encoding scheme*. This is performed to extract all the html tags and to get the pure text contents of the articles. Second, the IR system is constructed using the relational database model as explained above. This step involves tokenization, stop-word removal, root extraction, and term weighting.

4.2 Extracting the Root

In general, to extract Arabic roots from their words, the stemmer has to process each word in the following order [Khoja, 1999]:

- Removing the Definite Article ال "al"
- Removing the Conjunction Letter و "w"
- Removing Suffixes
- Removing Prefixes
- Pattern Matching

The following example demonstrates the whole stemming process applied to the Arabic word وليدرسوها *"wlydrsooha",* which is mapped to the complete English sentence *"and they are going to study it"*. The root of this word can be extracted as follows:

(*w*)-(*l*)-(*y*)-*drs*-(*oo*)-(*ha*) (ها)(و)درس-(ي)-(ل)-(و)

1. Removing the conjunction letter (*w*) (و)
   ➔ (ها)(و)درس-(ي)-(ل)
2. Removing the suffix (*ha*) (ها), which indicates a feminine, singular patient
   ➔ (و)درس-(ي)-(ل)
3. Removing the suffix: (*oo*) (و), which indicates a masculine third person plural agent
   ➔ درس-(ي)-(ل)
4. Removing the preposition prefix (*l*) (ل)
   ➔ درس-(ي)
5. Removing the prefix: (*y*) (ي), which indicates a 3rd person, present tense ➔ درس
6. The pattern فعل *F9L* has the same length as the word درس *drs*. Then the stemmer detects that the word درس matches the pattern فعل, since all the letters of the word match those in the pattern (i.e. ف، ع، ل)
7. Finally, the stemmer checks the trilateral roots table and concludes that the root درس *drs* (*he studied*) is a valid root.

5 Question Processing in QARAB

Achieving question understanding requires deep semantic processing, which is a non-trivial task of natural language processing. In fact, Arabic NLP does not have solid research at the semantic level. Therefore, QARAB uses shallow language understanding to process questions and it does not attempt to understand the content of the question at a deep, semantic level.

QARAB treats the incoming question as a *"bag of words"* against which the index file is searched to obtain a list of ranked documents that possibly contain the answer. The question processing begins by performing tokenization to extract individual terms. Then, the stop-words are removed. The remaining words are tagged for part-of-speech in an attempt to highlight the main words that should appear in the hypothesized answer. The greatest effort should be spent on identifying proper names, as they are our best guidance to identify the possible answer. The interrogative particles that precede the questions will determine what types of answers are expected as shown in Table 1.

5.1 Query Expansion

To achieve better search and retrieval results the query is expanded to include all the terms (verbs and nouns derived from verbs) that occur in the index file and have the same roots, which were extracted from the original query words. The result of the query processing is then passed to the IR system to retrieve a ranked list of documents that match the terms of the query.

5.2 Query Type

Questions are classified based on a set of known *"question types"*. These question types help us to determine the type of processing needed to identify and extract the final answer. The QARAB system recognizes the following set of question types (Table1):

Table 1. Question Types Processed by the QARAB System

Query Starting with		Query Type
من	Who, Whose	Person
متى	When	Date, Time
ماء، ماذا	What, Which	Organization, Product, Event
اين	Where	Location (natural, political)
كم	How Much, How Many	Number, Quantity

There are two other types of question particles, namely كيف and لماذا (How and Why). Although they will form legitimate query structures, they require long and procedural answers and are beyond the scope of our research. It is worth mentioning that the How and the Why queries also caused problems for many TREC-8 participants.

5.3 Query Keyword Identification

The remaining words of the query (after removing punctuation and stop-words) are tagged for part of speech. This process requires using the *Type-Finder* & the *Proper Name-Finder* system implemented by Abuleil [1999]. Verbs, which almost always follow clear morphological patterns, are the easiest to identify. Nouns, especially proper nouns, are considered as our best guide to find the expected answer from the relevant documents returned by the IR system. They have to occur within the selected answer passage and must be in the same order as they appeared in the original question. A list of keywords to identify *personal names*, *organization names*, *locations*, *numbers*, *money* and *dates*, has been constructed for Arabic to help in identifying proper names.

6 Answer Processing in QARAB

The input to the QARAB *Answer Generator* module is a natural language question and a small set of ranked documents. The question is first processed by tagging all the words. Then the set of relevant documents that may contain the answer are retrieved by the IR system. In the answer generation process, the passages of the relevant documents that match (are similar to) the query's *"bag of words"* closely are collected for further processing. The answer zones usually include most of the terms appearing in the original query in addition to the proper nouns that should appear in the final answer. The following example illustrates the whole process taken by the QARAB system to answer a question.

The following document extracted from the newspaper *Al-Raya* published in Qatar was processed by the IR system:

قال محافظ البنك المركزي الكويتي الشيخ سالم عبد العزيز الصباح امس ان بلاده ليس لديها النية لخفض قيمة الدينار الكويتي للحد من العجز المتزايد في الميزانية . وقال بأن خفض قيمة الدينار سيضر باقتصاد الكويت ومصداقيتها في الاسواق المالية الدولية.

وأكد الشيخ سالم ان البنك المركزي لن يخفض قيمة العملة كوسيلة لتقليص العجز في الميزانية . ومن المتوقع ان يبلغ العجز في ميزانية عام ١٩٩٨/١٩٩٩ التي تنتهي في يونيو ستة مليارات دولار .

Translated by ajeeb: www.ajeeb.com

Said the governor of the Kuwaiti central bank is sheikh Salem Abd Al-Aziz Al-Sabah yesterday that his countries not have her the intention to the Kuwaiti dinar devaluation to the restriction from the increasing inability in the budget. And believed that the dinar devaluation will harm the Kuwait economy and her credibility in the international exchanges.
And confirmed the sheikh Salem is that the central bank will not reduce the currency value as a means to the inability reduction in the budget. From it is expected that the inability in a budget reaches a year 1998 / 1999 that ends in June is six billions dollar.

Assume the user posed the following question to QARAB:

من هو محافظ البنك المركزي الكويتي والذي قال بأن بلاده ليس لديها النية لخفض قيمة الدينار للحد من عجز الميزانية؟

Translated by ajeeb: www.ajeeb.com

Who he is the governor of the Kuwaiti central bank and that believed by that his country not have her the intention to the dinar devaluation to the restriction from the budget inability?

Step 1: The query is processed as shown in Table 2

Table 2. Query Processing

Token		Stem	Part of Speech	Stop Word
هو	he	هو	Pronoun	Yes
محافظ	governor	محافظ	Noun	
البنك	bank	بنك	Noun	
المركزي	central	مركز	Noun	
الكويتي	Kuwaiti	كويت	Noun	
و	and	و	Conjunction	Yes
الذي	that	الذي	Pronoun	Yes
قال	said	قال	Verb	
بأن	that	بأن	Particle	Yes
بلاده	his country	بلاد	Noun	
ليس	not	ليس	Verb	Yes
لديها	have	لدى	Particle	Yes
النية	intention	نية	Noun	
لخفض	devaluation	خفض	Noun	
قيمة	value	قيمة	Noun	
الدينار	dinar	دينار	Noun	
للحد	restriction	حد	Noun	
من	from	من	Preposition	Yes
عجز	inability	عجز	Noun	
الميزانية	budget	ميزانية	Noun	
؟	؟	؟	Punctuation	Yes

Step 2: QARAB constructs the query as a "*bag of words*" and passes it to the IR system

Table 3. *Bag of words*

محافظ
بنك
مركز
كويت
بلاد
نية
خفض
قيمة
دينار
حد
عجز
ميزانية

Assume the system returned the following document as the top ranked document that closely matches the query.

قال **محافظ البنك المركزي الكويتي** الشيخ سالم عبد العزيز الصباح امس ان **بلاده** ليس لديها **النية لخفض قيمة الدينار الكويتي للحد** من **العجز** المتزايد في **الميزانية**. وقال بأن **خفض قيمة الدينار** سيضر باقتصاد **الكويت** ومصداقيتها في الاسواق المالية الدولية.
وأكد الشيخ سالم ان **البنك المركزي** لن **يخفض قيمة** العملة كوسيلة لتقليص **العجز** في **الميزانية** . ومن المتوقع ان يبلغ **العجز** في **ميزانية** عام ١٩٩٨/١٩٩٩ التي تنتهي في يونيو ستة مليارات دولار.

Step 3: Determine the expected type of the answer

من "*Who*" ➔ Person Name

Step 4: Generating the answer

The Answer Generator looks for keywords that might identify a person name using the personal names keywords. The input to the Answer Generator is the "bag of words" and the paragraphs extracted from the top ranked relevant documents.

قال **محافظ البنك المركزي الكويتي** الشيخ سالم عبد العزيز الصباح امس ان **بلاده** ليس لديها **النية لخفض قيمة الدينار الكويتي للحد** من **العجز** المتزايد في **الميزانية**. وقال بأن **خفض قيمة الدينار** سيضر باقتصاد **الكويت** ومصداقيتها في الاسواق المالية الدولية.

وأكد الشيخ سالم ان **البنك المركزي** لن **يخفض قيمة** العملة كوسيلة لتقليص **العجز** في **الميزانية.** ومن المتوقع ان يبلغ **العجز** في **ميزانية** عام ١٩٩٨/١٩٩٩ التي تنتهي في يونيو ستة مليارات دولار.

Keywords that might identify personal names:

The keyword **الشيخ** *sheikh* is used to mark an Arabic personal name.
The keyword **عبد** *Abd* is used to mark the beginning of a personal name.

قال محافظ البنك المركزي الكويتي **الشيخ** سالم **عبد** العزيز الصباح امس ان بلاده ليس لديها النية لخفض قيمة الدينار الكويتي للحد من العجز المتزايد في الميزانية. وقال بأن خفض قيمة الدينار سيضر باقتصاد الكويت ومصداقيتها في الاسواق المالية الدولية .

وأكد **الشيخ** سالم ان البنك المركزي لن يخفض قيمة العملة كوسيلة لتقليص العجز في الميزانية . ومن المتوقع ان يبلغ العجز في ميزانية عام ١٩٩٨/١٩٩٩ التي تنتهي في يونيو ستة مليارات دولار .

The first paragraph has most of the query words and the keywords that might identify a personal name. Therefore, the first paragraph is returned as the potential answer.

قال محافظ البنك المركزي الكويتي **الشيخ سالم عبد العزيز الصباح** امس ان بلاده ليس لديها النية لخفض قيمة الدينار الكويتي للحد من العجز المتزايد في الميزانية. وقال بأن خفض قيمة الدينار سيضر باقتصاد الكويت ومصداقيتها في الاسواق المالية الدولية .

7 Conclusion

We have described an approach to question answering system that provides short answers to questions expressed in the Arabic language. The system utilizes techniques from IR and NLP to process a collection of Arabic text documents as its primary source of knowledge. An actual system named QARAB is implemented and an initial ad-hoc analysis seems to be promising. The overall success of the system is limited to the amount of available tools developed for the Arabic language. Work is undergoing to get retrieval integrated into the system and to extend the functionality of the NLP system by developing more sophisticated algorithms to produce a concise answer in a timely manner.

References

Abuleil, S., and Evens, M., 1998. "Discovering Lexical Information by Tagging Arabic Newspaper Text", Workshop on Semantic Language Processing. COLING-ACL '98, University of Montreal, Montreal, PQ, Canada, Aug. 16 1998, pp. 1-7.

Al-Daimi, K., and Abdel-Amir, M. 1994. "The Syntactic Analysis of Arabic by Machine". Computers and Humanities, Vol. 28, No. 1, pp. 29-37.

Allan, J., Callan, J., Feng, F-F., and Malin D. 1999. "INQUERY and TREC-8". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 637-645.

Ask Jeeves. 1996. www.ask.com Site last visited in March 2001.

Breck, E., Burger, J., Ferro, L., House, D., Light, M., and Mani, I. 1999. "A Sys Called Qanda". Proceedings of the 8th Text REtrieval Conference, NIST Special Publications, pp. 499-507.

Budzik, J. and Hammond, K. 1999. "Q&A: A System for the Capture, Organization and Reuse of Expertise". Proceedings of the Sixty-second Annual Meeting of the American Society for Information Science. Information

Today, Inc., Medford, NJ. Available on the Web at http://dent.infolab.nwu.edu/infolab/downloads/papers/paper10061.pdf. Site last visited in August 2001.

Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N., and Schoenberg, S. 1997. "Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System". AI Magazine, Vol. 18, No.2, pp. 57-66.

Cardie, C., Ng, V., Pierce, D., and Buckley, C. 2000. "Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System". Proceedings of the Sixth Applied Natural Language Processing Conference, pp. 180-187.

Cormack, G., Clarke, C., and Kisman, D. 1999. "Fast Automatic Passage Ranking (MultiText Experiments for TREC-8)". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 735-743.

Ferret, O., Grau, B., Illouz, G., Jacquemin, C., and Masson, N. 1999. "QALC - the Question-Answering Program of the Language and Cognition Group at LIMSI-CNRS". Proceedings of the 8th Text REtrieval Conference, NIST Special Publications, pp. 465-475.

Harabagiu, S., Pasca, M., and Maiorano, S. 2000. "Experiments with Open-Domain Textual Question Answering". Proceedings of 18th International Conference on Computational Linguistics (COLING-2000), Saarbrucken, Germany, pp. 292-298

Hull, D. 1999. "Xerox TREC-8 Question Answering Track Report". Proceedings of the 8thText REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 743-751.

Humphreys, K., Gaizauskas, R., Hepple, M., and Sanderson, M. 1999. "University of Sheffield TREC-8 Q & A System". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 707-717.

Jacobs, P., and Rau, L. 1990. "SCISOR: Extracting Information from On-line News". Communications of the ACM, Vol. 33, No.11, pp. 88-97.

Katz, B. 1997. "From Sentence Processing to Information Access on the World Wide Web". Proceedings of the American Association for Artificial Intelligence Conference, Spring Symposium, NLP for WWW, pp. 77-86.

Khoja, S. 1999. "Stemming Arabic Text". Available on the Web at: http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps. Site last visited in March 2001.

Kupiec, J. 1993. "MURAX: A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia". Proceedings of the 16th Annual Int. ACM SIGIR Conference, pp. 181-190.

Lehnert, W. 1978. The Process of Question Answering. Lawrence Erlbaum Associates, Hillsdale, NJ.

Lin, C-J, and Chen, H-H. 1999. "Description of Preliminary Results to TREC-8 QA Task". Proceedings of the 8th Text REtrieval Conference(TERC-8), NIST Special Publications 500-246, pp. 507-513.

Litkowski, K. 1999. "Question-Answering Using Semantic Relation Triples". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-248, pp. 349-357

Lundquist, C., Grossman, D., and Frieder, O. 1999. "Improving Relevance Feedback in the Vector Space Model". Proceedings of 6th ACM Annual Conference on Information and Knowledge Management (CIKM), pp. 16-23.

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., and Rus, V. 2000. "The Structure and Performance of an Open-Domain Question-Answering System". Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 563-570.

Oard, D., Wang, J., Lin, D., and Soboroff, I. 1999. "TREC-8 Experiments at Maryland: CLIR, QA and Routing". Proceedings of the 8th Text

REtrieval Conference (TERC-8), NIST Special Publications 500-246, pp. 623-637.

Ogden, B., Cowie, J., Ludovik, E. Molina-Salgado, H., Nirenburg, S., Sharples, N., and Sheremtyeva, S. 1999. "CRL's TREC-8 Systems Cross-Lingual IR, and Q&A". Proceedings of the 8th Text REtrieval Conference (TERC-8), NIST Special Publications 500-246, pp. 513-523.

Salton, G. 1971. The SMART Retrieval System Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, NJ.

Schank, R., and Abelson, R. 1977. Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum Associates, Hillsdale, NJ.

Shin, D-H, Kim, Y-H, Kim, S., Eom, J-H, Shin, H-J, and Zhang B-T. 1999. "SCAI TREC-8 Experiments". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 583-591.

Singhal, A., Abney, S., Bacchiani, M., Collins, M., Hindle, D., and Pereira, F. 1999. "AT&T at TREC-8". Proceedings of the 8th Text REtrieval Conference, NIST Special Publications, pp. 317-331.

Srihari, R., and Li, W. 1999. "Information Extraction Supported Question Answering". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 185-197.

Takaki, T. 1999. "NTT DATA: Overview of System Approach at TREC-8 ad-hoc and Question Answering". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publications 500-246, pp. 523-531.

TREC-8. 1999. NIST Special Publication 500-246: The Eighth Text REtrieval Conference. Available on the Web at: http://trec.nist.gov/pubs/trec8/t8_proceedings.html. Site last visited in August 2001.

TREC-9. 2000. NIST Special Publication: The Ninth Text REtrieval Conference. Available on the Web at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html. Site last visited in August 2001.

Vicedo, J., and Ferrández, A. 2000. "Importance of Pronominal Anaphora Resolution in Question- Answering System". Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 555-562.

Voorhees, E., and Tice, D. 1999. "The TREC-8 Question Answering Track Evaluation". Proceedings of the 8th Text REtrieval Conference (TREC-8), NIST Special Publication 500-246, pp. 83-106.

Voorhees, E., and Tice, D. 2000. "Building a Question Answering Test Collection". Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 200-207.

Winograd, T. 1972. Understanding Natural Language. Academic Press, New York, NY.

Woods, W., Kaplan, R., and Webber, B. 1972. "The Lunar Sciences Natural Language Information System: Final Report". Bolt Beranek and Newman Inc. (BBN), Report No. 2378.