American Journal of Applied Sciences 6 (4): 797-805, 2009
ISSN 1546-9239
© 2009 Science Publications

A New Question Answering System for the Arabic Language

[1]Ghassan Kanaan, [2]Awni Hammouri, [1]Riyad Al-Shalabi and [3]Majdi Swalha
[1]Faculty of Information Systems and Technology
Arab Academy for Banking and Financial Science, P.O. Box 13190 Amman 11942 Jordan
[2]Departmentt of Information Technology, Faculty of Science, Mu'tah University, Mu'tah, Karak, 61710, Jordan
[3]School of Computing, University of Leeds, Leeds LS29JT UK

Abstract: We depict the architecture of a question answering system and methodically evaluate contributions of different system components to accuracy. The system differs from most question answering systems in its dependency on data redundancy rather than complicated linguistic analyses of either questions or contender answers. Because a wrong answer is often worse than no answer. A Question Answering (QA) system is a system that takes natural language questions expressed in the Arabic language then attempts to provide short answers. In order to handle this problem, traditional information retrieval techniques joined with a sophisticated natural language processing approach have been used in this research work. Using keyword matching, simple structures extracted from both the question and the candidate documents selected by the IR system were used in the process of identifying the answer. In order to perform this process, we used an existing tagger to identify proper names and other crucial lexical items and build lexical entries. Also provide an analysis of Arabic question forms and attempt to formulate better kinds of answers that users find more appropriate.

Key words: Question answering system, natural language processing, information retrieval

INTRODUCTION

There is a daily increase in the amount of data available on the Internet nowadays. Users surf the Internet to search for specific questions that they have in mind and hope to find short, precise answers. Users always prefer to express their questions in their native language without being restricted to a specific query language. For this reason, Question Answering Systems (QAS) are a necessity. Our QAS is the result of coupling a traditional Information Retrieval (IR) technique with a sophisticated Natural Language Processing (NLP) approach. The system 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 and then the NLP technique is used to parse the question, analyze the most appropriate documents returned by the IR system and formulate an answer.

Objectives of developing arabic QAS: The Arabic Language occupied the sixth rank of languages in the world with about 186 million native speakers. In addition, research in the Arabic Language is still narrow and faces some problems, thus there is a great need for a system that recognizes the Arabic language. Meeting such requirements is the main motive for developing our QAS. The new Arabic QAS introduced in this paper is a first step in our strategy to develop a complete and more advanced system. The development approach that has been adopted for developing the QAS is the spiral model, therefore the first iterative loop will deliver first release of this system.

The benefits of developing our QAS are include:

- Supporting the Arabic language worldwide
- Providing a system that is more effective, precise and easy to use compared to other existing systems
- Simplifying the complexity of preexisting algorithms

MATERIALS AND METHODS

Non-arabic question answering systems: Worldwide, most of the information retrieval research has been done on the English language. Several systems have been implemented and some of them have been improved as documented in the literature of information retrieval. Researchers have been attracted to develop an open-domain QA systems based on collections of real

Corresponding Author: Awni Hammouri, Department of Information Technology, Mu'tah University, Mu'tah, Karak, 61710, Jordan Tel: +962795756060

world documents, especially the World Wide Web. These systems seem to typically have long response times and/or accuracy rates that may not be acceptable to users[11]. In this study, two wide-spread existing systems have been described, as background to our contribution.

AnswerBus is an open-domain question answering system based on sentence level information retrieval. It accepts user's natural questions in English, German, Spanish, French, Italian and Portuguese and extracts possible answers from the Web. Five search engine and directories (Google, Yahoo, WiseNut, Alta Vista and Yahoo News) are used to retrieve Web pages that contain answers. AnswerBus then extracts sentences that are contain answers[11].

Another well-known Question-Answering System is AskMSR, unlike other systems, AskMSR depends on data redundancy rather than linguistic analysis of questions or candidate answers. This system also explores strategies for predicting when the question answering system is likely to give wrong answer[12].

Arabic question answering systems: Because of the increasing growth of Arabic textual data on the web and the improvement of Arabic software for browsing the web in the last decade, the need to implement a QA system that supports the Arabic language becomes an essential requirement. An existing example of such systems is QARAB.

QARAB, implemented by Bassam Hammo and others, provides short answers to questions expressed in the Arabic language[1].

The aim of the QARAB system is to identify text passages that answer natural language questions. The assumptions for the system can be summarized as follows:

- The answer exists in a collection of Arabic documents
- 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 the QARAB system is:

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

Arabic language structure: The Arabic language is the language of the Holy Quran. It is one of the six official languages of the United Nations and the mother tongue of approximately 300 million people. It can be classified into three types: Classical Arabic (العربية الفصحى), Modern Standard Arabic (العربية الحديثة) and Colloquial Arabic dialects (العربية العامية). Classical Arabic is fully vowelized and it is the language of the holy Quran. Modern Standard Arabic (MSA) is the official language throughout the Arab world. It is used in official documents, newspapers and magazines, in educational fields and for communication between Arabs of different nationalities. Colloquial Arabic dialects, on the other hand, are the languages spoken in the different Arab countries; the spoken forms of Arabic vary widely and each Arab country has its own dialect. Dialects are spoken in most informal settings, such as at home, with friends, or while shopping. MSA can be viewed as classical, since there have been no major changes modifying the structure of the classical language. The MSA, however, differs from Classical Arabic in two aspects: by adopting easy stylistic changes and by expanding the lexicon to include new technical terms. In addition, MSA has a rich morphology, based on consonantal roots, which depends on vowel changes and in some cases consonantal insertions and deletions to create inflections and derivations

The model of the Arabic word: An Arabic word is a string of characters from the Arabic alphabet that may or may not have diacritics Also it may be an original Arabic word, or it may be arabized. The original Arabic words are divided in turn into two subcategories; Derivative Arabic words, which are the verbs and nouns that are formed according to the Arabic derivation rules and Fixed Arabic words, which are a set of words molded by Arabs, in ancient times, that do not obey the Arabic derivation rules (Fig.1). The arabized words are nouns borrowed from foreign languages (perhaps with some phonetic adjustments to suit the Arabic pronunciation) that have become common among native Arabic speakers[13].

Arabic word categories: Arabic grammarians traditionally classify words into three main categories. These categories are also divided into subcategories, which collectively cover the whole of the Arabic language. These categories are:

Noun: A noun in Arabic is a name or a word that describes a person, thing, or an idea.

The linguistic attributes of nouns that have been used in this tag set are:

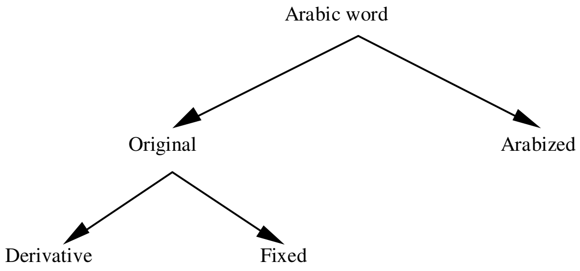


Fig. 1: The classification of the Arabic words

- **Gender:** Masculine
  Feminine Neuter
- **Number:** Singular
  Plural Dual
- **Person:** First
  Second Third
- **Case:** Nominative
  Accusative Genitive
- **Definiteness:** Definite
  Indefinite

Verbs: Verbs indicate an action, although the tenses and aspects are different. Verb categories are divided into subcategories such as Perfect, Imperfect and Imperative.

The verbal attributes that have been used in our tag set are:

Gender: Masculine
Feminine Neuter
Number: Singular
Plural
Dual
Person: First Second
Third

Particles: The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions and Interjections.

The subcategories of particle are:

- Prepositions -
Adverbial - Conjunctions
- Interjections - Exceptions
- Negatives
- Answers
- Explanations - Subordinates

Our approach: The objective of our project is to build a Question Answering System for the Arabic language. This automatic Question Answering system attempts to find answers for user queries. An automatic QA System for the Arabic language starts by asking the user to enter a natural language question and then attempts to find an answer on our Arabic QA System.

The goal of our research is to develop a model for answering Arabic questions using Arabic grammar and morph syntactic patterns. The user starts by entering Arabic question in a text box, then the system processes this question, in order to find appropriate answers. The user can see the processes that are operating and the progress that is being made by the system by clicking on the search button.

The computational approach consists of three parts: Inputs and Outputs of the system, System Architecture and Processes. Below is a detailed description for each part of the system (Fig.2).

Inputs and outputs: The input for the Question Answering system is a short question written in the Arabic language and a small set of ranked documents retrieved by the IR system. At this point our system cannot handle a short question that begins with the question word (ماذا, كيف) 'how' or 'why' because of the complex processing such questions need. The main output is a paragraph that contains the answer.

Processors and processes: The algorithm of the full automatic Arabic question answering system is made up of several processes and processors, each of which is responsible for one deterministic task.

First, comes the question analysis process; this is done by the NLP.

Second, the system retrieves candidate documents containing information relevant to the user's query; this is done by the IR System.

Third, the system chooses the most appropriate document according to the top similarity values, which are calculated by the IR system..

Fourth, the system generates the answer.

Description of the main algorithm of the Arabic QAS system:

```
Algorithm: Arabic_question_answering
Input: A natural Arabic language question.
Output: A set of ranked documents, with texts that are
highly likely to contain the answer highlighted.
Begin Arabic_question_answering
While UserQuestion is not empty
Tokenize the UserQuestion into Tokens;
    Determine the QuestionType;
    Determine the ProperNounPhrase;;
    Extract the Root of each NonStopWord;
    Convert UserQuestion into Query q;
    Index the Query q;
    Find IndexWeight of each Index term in the Query
    q;
End while;
While DocumentSet is not empty
    Calculate sim (dj, q);
End while;

Sort sim (dj, q) in a descending order;
Retrieve the first Documents;
Generate the Answer
End Arabic_question_answering;
```

Question analyzer: We use an existing Natural Language Processing System (NLP), which is called Tagger[14], to analyze both the user question and the documents, it determines the assignment each word with its root and the part of the speech, then saves them in the project database. This system composed of a set of tools to tokenize and tag Arabic text, identify some features of the tokens and identify proper names.

The NLP system comprises the following modules:

- The *Tokenizer*, which is used to extract the tokens from the both the query and the documents
- The Tagging System (or type finder), grammatical tagging (or part-of-speech tagging) is the process of assigning part of speech to words based on their context
- The Feature Finder, which is used to determine the features of each word
- The Proper Noun Phrase Parser, which is used to tag proper nouns

The Tagger was designed to construct an Arabic lexicon. The *Lexicon* is a collection of representations for words used by a linguistic processor as a source of word specific information; this representation contains information on the morphology, phonology, syntactic argument structure and semantics of the word[5].

How the NLP (part of speech tagger) works?: First comes the tokenizing process, the stemming process, the affix extractor process and the pattern recognizer, which are applied in the first part of the tagging process.

Second, the lexical word matcher is responsible for assigning tags to the words in the document that match words stored previously in the lexicon database.

Third, the noun tagger process, which is responsible for applying linguistic rules for nouns to the document's words, if one of the rules matches for any word in the document, then its tag will be assigned and stored.

Fourth, the verb tagger process, which is responsible for applying linguistic rules for verbs to the document's words, if one of these rules matches for any of the untagged words, its tag will be assigned and stored[14].

Retrieve candidate documents: IR deals with the representation, storage, organization of and accesses to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested[4].

The main object of the IR module is to retrieve all relevant documents for a user query and only those relevant documents.

The IR system that we use is implemented from the scratch based on Salton's vector space model. And here is a brief description of the model:

Vector space model: This model defines a vector that represents each document and a vector that represents the query[7]. The model works by assigning weights to index terms in queries and documents. We use this weight to compute the degree of similarity between each document and the user query.

The degree of similarity of the document dj with regard to the query q is given by the cosine of the angle between the vectors dj and q[4].

$$\text{Sim (dj,q)} = \vec{\text{dj}} \,.\, \vec{\text{q}} \,/\, |\vec{\text{dj}}| \times |\vec{\text{q}}|$$

$$\text{Sim(dj, q)} = \frac{\sum \text{i} = 1^{\text{t}} \text{Wi.j} \times \text{wi, q}}{\sqrt{\sum {}^{\text{t}}\text{i} = \text{w}^{2}{}_{\text{i, j}}} \sqrt{\sum {}^{\text{t}}\text{i} = \text{w}^{2}{}_{\text{i,q}}}}$$

Where:

Sim(dj,q) = Similarity between document dj and query q

t = Number of index terms in the system

Wi,j = Weight of the index term i in the document j

Wi,q = Weight of the index i in the query q

to calculate the weight of each term we use the following formula:-

$$\text{w}_{\text{i,j}} = \text{fi,j*idfi}$$

Where: fi,j is called the normalized frequency and calculating according to the following formula:

$$\text{fi,j} = \text{freq}_{\text{i,j}}/\text{Max}_{\text{I}}\ \text{freq}_{\text{i,j}}$$

freqi,j = Frequency of the term ki in the document j
$\text{Max}_l\text{freq}_{i,j}$ = Maximum frequency computed in the document j

Idfi is called the inverse document frequency for the index term i, idfi is given by:

$$\text{Idfi} = \log(N/n_i)$$

where:
N = Total number of document in the system
n_i = Number of documents in which the index term k_i appears

The main reason for using this model is that the answer set of the retrieved ranked document is more precise than other IR models. The purpose of the IR system is to search and then retrieve candidate documents containing information relevant to the user's query.

Implementing the information retrieval system: We implement the IR system using a relational database management system (RDBMS) along with Microsoft Access 2000 tools. It contains the following database relations:

Words relation: is the main relation of the IR system. It is responsible for storing all document words; every process references it to get information needed to retrieve the relevant documents (Fig. 3).

Query Weight relation: This relation of the project database is responsible for storing the weight of the query words (Fig. 4).

Similarity of the query relation: This relation stores the cosine similarity between the documents and the user query (Fig. 5).

The basic outline of processing in the IR system: 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.

Saving the Roots Extracted from the Tagger (the stemming process): This process is an essential part of our program, since many processes to follow use its output. It is responsible for extracting the roots of all words in the document; the stemming process extracts roots constructed of three letters and stores the Root in the attribute of the project WORD relation to enable referencing it when it is needed.

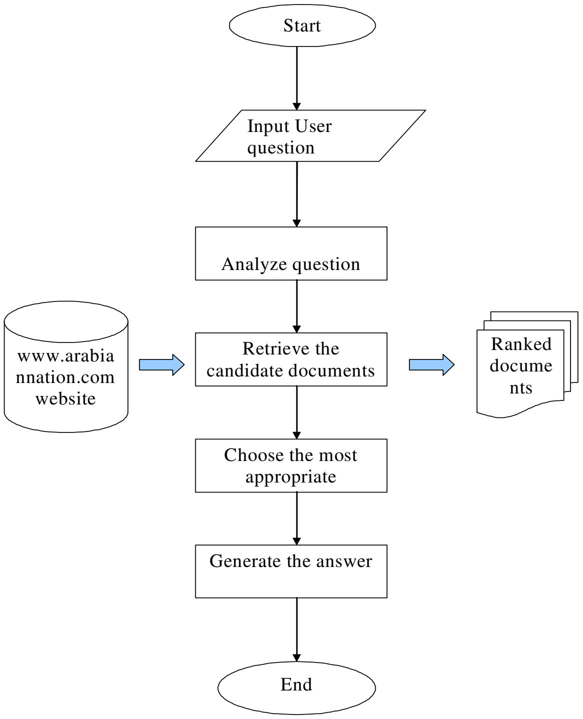
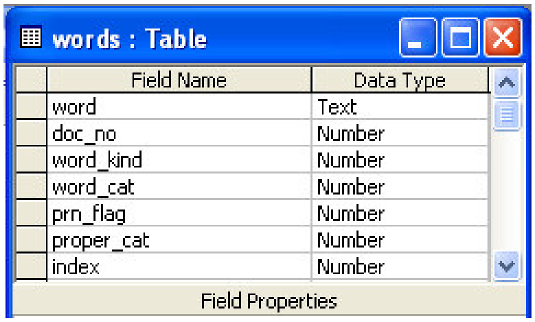
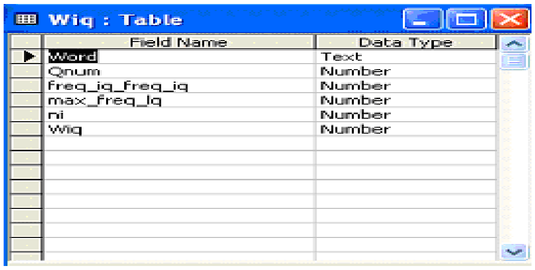

Start
Input User question
Analyze question
www.arabiannation.com website
Retrieve the candidate documents
Ranked documents
Choose the most appropriate
Generate the answer
End

Fig. 2: The flow chart of the QAS system



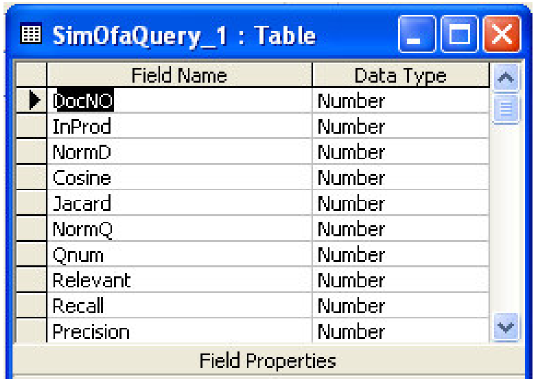
Field Name	Data Type
word	Text
doc_no	Number
word_kind	Number
word_cat	Number
prn_flag	Number
proper_cat	Number
index	Number

Fig. 3: The words relation



Field Name	Data Type
Word	Text
Qnum	Number
freq_iq_freq_iq	Number
max_freq_iq	Number
ni	Number
Wiq	Number

Fig. 4: Weight of a query relation



SimOfaQuery_1 : Table

Field Name	Data Type
DocNO	Number
InProd	Number
NormD	Number
Cosine	Number
Jacard	Number
NormQ	Number
Qnum	Number
Relevant	Number
Recall	Number
Precision	Number

Field Properties

Fig. 5: Similarity of query relation

In general, to extract Arabic roots from their words, the stemmer has to process each word in the following order[9]:

- Removing the Definite Article “al “"ل ا”
- Removing the Conjunction Letter و “w”
- Removing Suffixes
- Removing Prefixes
- Pattern Matching

How the QA system works: The work of our system contains two main steps: Question processing and answer processing; here is a complete explanation of each step:

The question processing step: In our approach we will use approximately the same method used by Hammo *et al.* in their approach in QARAB[1].

Because of the lack of research at the semantic level of the Arabic NLP system; the system will be somewhat shallow in understanding and processing the user's question, but it doesn't matter how deeply the question is understood.

First, the question is tokenized to extract individual terms. Then, the stop-words are removed. After these steps the question will be treated as a "bag of words" used to search the index file and obtain a list of ranked documents for a short paragraph that may contain the answer. Finally we identify proper names as they are our best guides to identify the possible answers.

Here is a detailed analysis of each process:

Query type: The particles that precede the question will determine the type of answer. The system will recognize the following question set:

Question Starting with	Question Class	
من	Who, Whose	Person, Group
متى	When	Date, Time
ماذا	What	Organization, Product, Event
اين	Where	Location
كم	How Much, How Many	Number, Quantity

Two other types of question particles, namely كيف and لماذا (How and why), are beyond the scope of our approach because they require long and procedural answers.

Generating the answer: The input to the answer-generation component of our system is a natural language question and a small set of ranked documents. First the question is processed by tokenizing 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 closely similar to) the query’s bag of words are collected for further processing. The answer zones usually include most of the terms appearing in the original query in addition to proper nouns that should appear in the final answer.

RESULTS AND DISCUSSION

Retrieval performance evaluation: Any Information Retrieval System should show how precise is the answer set.

If the user query is vague, the request will retrieve documents that do not contain the exact answer and must be ranked according to their relevance to the query.

Each retrieval performance evaluation for an IR System must be based on a collection of documents, a set of example information requests and a set of relevant documents. This collection is called the test reference collection and evaluation measure[4].

Our test reference collection consists of 25 documents gathered from the Internet, 12 queries (questions) and some relevant documents provided by us.

The strategy (Vector Space Model) will retrieve the documents that maybe relevant to the query; then we depend on the two most used retrieval measures to evaluate of the effectiveness of an information retrieval system, Recall and Precession. Suppose R is the set of relevant documents for the given query[4].

- |R| is the number of documents in the set R
- A is the set of retrieved documents for the query q
- |A| is the number of documents in the set A

- |Ra| is the number of documents in the intersection of the set R and A

Recall is the fraction of the relevant documents that have been retrieved.

$$\text{Recall} = \frac{|Ra|}{|R|}$$

Precision is the fraction of the retrieved documents that are relevant.

$$\text{Precision} = \frac{|Ra|}{|A|}$$

Since, Recall and Precision are calculated for one query at a time and there is a distinct recall level for each queryan interpolation procedure is necessary. We used an interpolating procedure with eleven standard recall levels.

Let Rj, j ε {0,1,2,…,10}, be the relevance at the jth standard recall level [e.g., r4 is the relevance at the recall level 40% . Then

$$P(rj) = \max rj <= r <= rj+1 \; P(r)$$

Which states that the interpolated precision at jth standard recall level is the maximum known precision at any recall level between the jth recall level and the (j+1)th recall level[4].

Figure 5 shows the average interpolated recall and precision after applying the steps described above.

The diagram shows that at the recall levels 0, 10 and 20%, the interpolated precision is equal to 100% and at recall levels 90 and 100% it is equal to 43%.

These results reflect the performance of our Information Retrieval system, which is close to the reported performance of the traditional Vector Space model.

Implementation: We built a program that applies all rules described previously using Microsoft Visual Basic 6.0 and Microsoft Access Database. The figures below shows the program forms used to input text and show the results of applying rules that construct the IR system for Arabic language automatically.

The first form in Fig. 6 shows the main screen of our system. It contains the program name and three buttons labeled “Enter”(1), “All Word”(2), “End “(3).

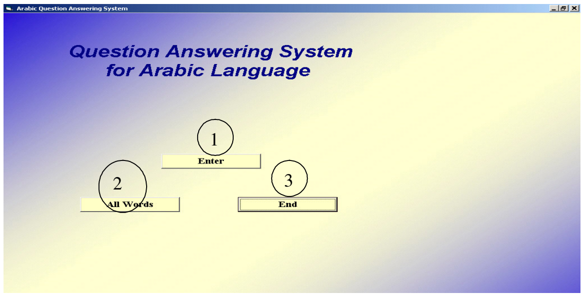
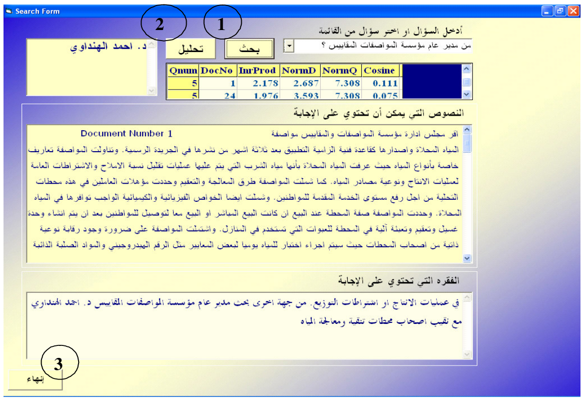


Fig. 6: Program First form



أدخل السؤال او اختر سؤال من القائمة

من مدير عام مؤسسة المواصفات المقاييس ؟

د. احمد الهنداوي

Qnum	DocNo	InrProd	NormD	NormQ	Cosine
5	1	2.178	2.687	7.308	0.111
5	24	1.976	3.593	7.308	0.075

النصوص التي يمكن أن تحتوي على الإجابة

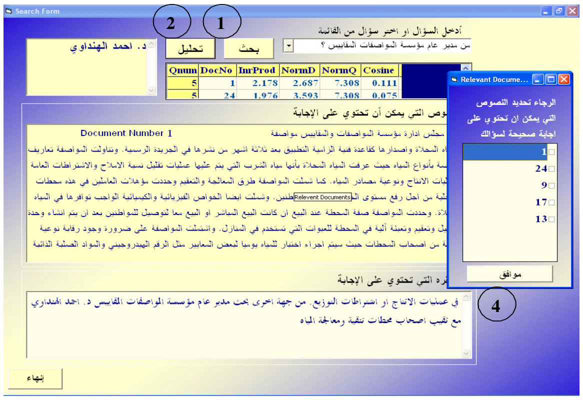
Document Number 1

اقر مجلس ادارة مؤسسة المواصفات والمقاييس مواصفة المياه المحلاة واصدارها كقاعدة فنية الزامية التطبيق بعد ثلاثة اشهر من نشرها في الجريدة الرسمية. وتناولت المواصفة تعاريف خاصة بأنواع المياه حيث عرفت المياه المحلاة بأنها مياه الشرب التي يتم عليها عمليات تقليل نسبة الاملاح والاشتراطات العامة لعمليات الانتاج ونوعية مصادر المياه. كما شملت المواصفة طرق المعالجة والتعقيم وحددت مؤهلات العاملين في هذه محطات التحلية من اجل رفع مستوى الخدمة المقدمة للمواطنين. وشملت ايضا الخواص الفيزيائية والكيميائية الواجب توافرها في المياه المحلاة. وحددت المواصفة صفة المحطة عند البيع ان كانت البيع المباشر او البيع معا لتوصيل للمواطنين بعد ان يتم انشاء وحدة غسيل وتعقيم وتعبئة آلية في المحطة للعبوات التي تستخدم في المنازل. واشتملت المواصفة على ضرورة وجود رقابة نوعية ذاتية من اصحاب المحطات حيث سيتم اجراء اختبار للمياه يوميا لبعض المعايير مثل الرقم الهيدروجيني والمواد الصلبة الذائبة

الفقره التي تحتوي على الإجابة

في عمليات الانتاج او اشتراطات التوزيع. من جهة اخرى بحث مدير عام مؤسسة المواصفات المقاييس د. احمد الهنداوي مع نقيب اصحاب محطات تنقية ومعالجة المياه

Fig. 7: The second form



الرجاء تحديد النصوص التي يمكن ان تحتوي على اجابة صحيحة لسؤالك

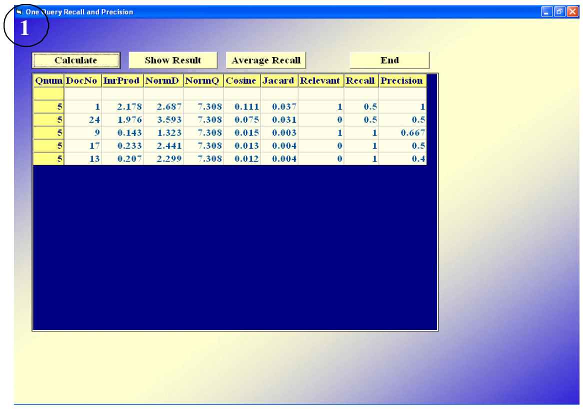
□ 1
□ 24
□ 9
□ 17
□ 13

موافق

Fig. 8: The third form

The second form is shown in Fig. 7, which is invoked by pressing the button labeled “Enter” (1) in the first form. It is responsible for searching for answers to the user question, that has been entered in, or selected from list box .The answer to the user's question will be displayed in the text box by pressing on the button labeled "بحث"(1). The button labeled "End"(3) terminates the program.

The third form is shown in Fig. 8. It will be displayed when the user presses on the button labeled "تحليل" (2). As a result, a dialog box that shows the



Qnun	DocNo	InrProd	NormD	NormQ	Cosine	Jacard	Relevant	Recall	Precision
5	1	2.178	2.687	7.308	0.111	0.037	1	0.5	1
5	24	1.976	3.593	7.308	0.075	0.031	0	0.5	0.5
5	9	0.143	1.323	7.308	0.015	0.003	1	1	0.667
5	17	0.233	2.441	7.308	0.013	0.004	0	1	0.5
5	13	0.207	2.299	7.308	0.012	0.004	0	1	0.4

Fig. 9: The fourth form

number of all documents containing relevant information to user's question will be displayed to allow the user to choose the most appropriate document that contains the answer to the given question. The user press “موافق” (4) to go to the next form.

The fourth form is in Fig. 9. It is invoked by pressing the button labeled "موافق" (4) in the third form shown in Fig. 8. It is responsible for calculating the recall and precision measures and getting the relevant documents (1) "calculate".

CONCLUSION

We have described an approach to the construction of a question answering system that provides short answers to questions expressed in the Arabic language. The system utilizes techniques from Information Retrieval and Natural Language Processing to process a collection of Arabic text documents as its primary source of knowledge. The overall success of the system is limited by the number of available tools developed for the Arabic language. Work is undergoing to get retrieval integrated into the system and to extent the functionality of the NLP system by developing more sophisticated algorithms to produce a concise answer in a timely manner.

Future research: Our system deals with questions phrased in the natural language, which enables us to take advantage of IR techniques for Arabic documents, but sometimes users can't find the answer to the user questions in these Arabic documents. This is because these documents do not have the answer, so to benefit more from this system we can search in other documents in the English Language, starting with an Arabic question. The cross-language question/answering system given a Natural Language query in one language (say Arabic) finds answers for that query in textual documents written in another language (say English) and eventually expresses the answers found in the query language (Arabic). In contrast to a standard cross-language IR system, the NL queries are usually well-formed NL-query clauses (instead of a set of keywords).

ACKNOWLEDGMENT

The authors would like to thank Professor Martha Evens from Illinois Institute of Technology for her helpful comments and suggestions that reflected many improvements in the presentation of this study.

REFERENCES

1. Hammo, B., H. Abu-Salem, S. Lytinen and M. Evens, 2002. QARAB: A question answering system to support the Arabic language. Proceedings of the 40th Association for Computational Linguistics on Computational Approaches to Semetic Languages, ACL’02 University of Pennsylvania, PA, USA, 55-65.
2. Hammo, B., H. Abu-Salem, S. Lytinen and S. Abuleil, 2002. Identifying proper nouns for an Arabic question answering system. Proceedings of the 13th Midwest Artificial Intelligence and Cognitive Science Conference MAICS, Chicago, IL, USA, pp: 130-136.
3. Hammo, B., S. Ableil, S. Lytinen and M. Evens, 2004. Experimenting with a question answering system for the Arabic language. Comput. Humanities, 38: 397-415.
4. Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. Addison Wesley, Reading, MA, USA.
5. Tiedemann, J., 1997. Automatic lexicon extraction from aligned bilingual corpora. Diploma Thesis, Department of linguistics/ computational linguistics, University of Uppsala, Uppsala, Sweden.
6. Salton, G., 1971. The SMART Retrieval System-Experiments in Automatic Document Processing. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
7. Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM., 18: 613-620.
8. Chowdhury, A. and C.M. McCabe, 1998. Improving information retrieval systems using part of speech tagging. Technical Report, ISR, TR 1998-48.

9. Khoja, S. and R. Garside, 2001. Automatic tagging of an Arabic Corpus using APT. Arabic Linguistic Symposium, Salt Lake City, Utah.
10. Abuleil, S. and M. Evens, 1998. Discovering lexical information by tagging Arabic newspaper text. Workshop on Semitic Language Processing. Coling-ACL '98, pp: 1-7.
11. Zheng, Z., 2002. AnswerBus question answering system. Proceedings of the 2nd International Conference on Human Language Technology Research, pp: 399-404.
12. Brill, E., S. Dumais and M. Banko, 2002. An analysis of the AskMSR question-answering system. Proc. ACL-02 Conf. Empirical Methods Natr. Language Process., 10: 257-264.
13. Elaraby, M.A., 2000. A large-scale computational processor of the Arabic morphology and applications. Master's Thesis, Cairo University, Egypt.
14. Kanaan, G., R. Al-Shalabi and M. Swalha, 2003. Full automatic Arabic text tagging system. The Proceedings of the International Conference on Information Technology and Natural Sciences, pp: 258-267.