Toward a New Arabic Question Answering System

Imane Lahbari, Said El Alaoui, and Khalid Zidani
Computer Science and Modeling Laboratory, Sidi Mohammed Ben Abdellah University, Morocco

Abstract: *Question Answering Systems (QAS) aim at returning precise answers to user's questions that are written in natural language. In this paper, we describe our question processing and document retrieval as two components of Arabic QAS. First, we present Arabic question classification method based on SVM classifier and Li and Roth's [24] taxonomy. Then, we describe our proposed technique to transform an Arabic question, to a query which is available to get information from the Arabic Wikipedia. In this paper, we use a hybrid Arabic Part-of-Speech (POS) tagging and Arabic WordNet (AWN) for query expansion. We have conducted several experiments using Text Retrieval Conference (TREC) and Cross Lingual Evaluation Forum (CLEF) datasets. The obtained results have shown that the proposed method is more effective as compared with the existing methods.*

Keywords: *Natural language processing, Arabic question answering system, question classification, taxonomy, machine learning approach, SVM, decision-tree, naive bayes, POS tagging, query expansion, AWN.*

Received February 15, 2018; accepted April 18, 2018

1. Introduction

An Information Retrieval System (IRS) extracts relevant documents for a user's need. Usually, this need is expressed by a list of keywords. In many cases, classical search engines may not satisfy users' needs regarding the huge size of the available documents and the eventual irrelevance of their results. This issue often results in the need of human intervention and feedback in order to retrieve the requested information which is a waste of time and accuracy. To resolve this problem several Question Answering Systems (QAS) have been proposed.

A QAS should return an answer to a user question written in natural language. Its most important goal is to provide an effective answer efficiently and expeditiously. Such systems combine two important research domains: Information retrieval and Natural Language Processing (NLP). The first NLP project was an English-Russian automatic translator, built in 1954 at Georgetown University (Washington, USA). It was able to handle 250 words and six grammar rules. Therefore, NLP has other applications like Named Entity Recognition, Part of Speech Tagging, Summarizing, Information Retrieval, QAS and so on.

The Arabic language is one of the widely spoken languages in the world with more than 420 million speakers. It's a Semitic language and it is one of the six official languages of the United Nations.

We should emphasize that Arabic morphology is different than the Latin one and it's richer. Arabic is a derivational language and its vocabulary contains about 10000 roots. In this regard, this language requires specific pre-processing for NLP tasks. This language can be classified into three types: Classical Arabic (العربية الفصحى), Modern Arabic (العربية الفصحى الحديثة) and Colloquial Arabic (العربية العامية). Classical Arabic is a sophisticated language, that is, its terms are not easily understood by a simple listener. It's the language of the Holy Quran (Muslims sacred book). Modern Arabic respects all grammatical rules of the Classical Arabic, but with simple terms. It's the official language throughout the Arab world. Classical Arabic and modern Arabic may contain diacritics, which are a short vowel mark. The main purpose of diacritics is to provide a phonetic guide or a phonetic aid.

Colloquial Arabic depends, to some extent, upon the dialects spoken in each region. The Arabic dialects are spoken in informal settings.

Figure 1 presents the flowchart of QAS with 3 blocs:

- Question preprocessing.
- Information retrieval.
- Answer Processing.

The Arabic Question classification plays a vital role in QAS, its influences, positively or negatively, the whole system, because its results will be used by the other components. In this work, we propose a rule-based method to classify Arabic questions. We propose a set of rules to classify questions according to two taxonomies: Arabic taxonomy and Li and Roth [24] taxonomy.

This paper is organized as follows: section 2 presents the related work. Section 3 is devoted to the question classification. Paragraph 4 describes the process to get an extended query from a question and we present the search engine in section 5. Experiments are presented and discussed in section 6 and we conclude in section 7 with discussing the future works.

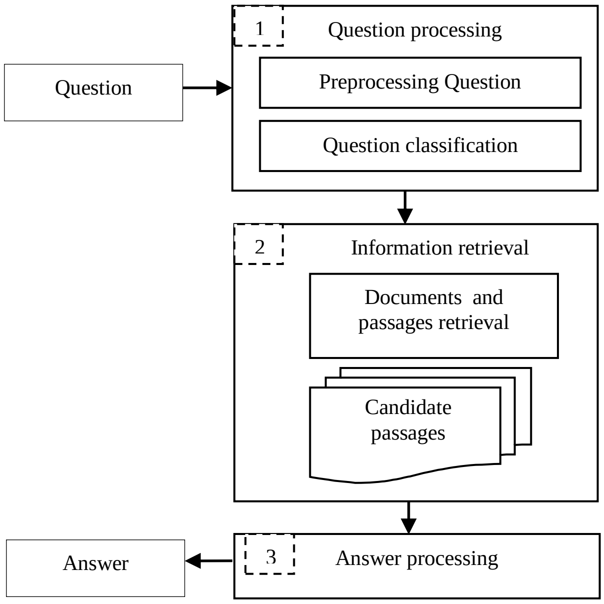

Figure 1. QAS architecture.

2. Related Work

There are three approaches to classify questions: rule-based approach, machine learning approach and hybrid approach, that combines rule-based approach and learning based approach. The first one is defined by specific rules depending on patterns. The second, machine learning approach, allows to classify questions after a learning step which needs an annotated data set.

In this work we have adopted the machine learning approach because it deals with all possible question types; it is flexible for the new data and less complicated than the rule-based approach.

Many algorithms have been applied to Text Classification. Most studies have been devoted to English and other Latin languages. However, very few researches have been carried out on Arabic text:

- El Kourdi *et al.* [14] classified Arabic web documents automatically using Naive Bayes (NB) which is a machine learning algorithm. Cross validation experiments were used to evaluate the obtained results. The categorization accuracy varies from one category to another with an average accuracy over all categories of 68.78 %.
- Maximum entropy (ME) used by El-Halees [13] and Sawaf [26] for classifying Arabic news articles. The classification accuracy was 80.41% and 62.7% by Sawaf without any morphological analysis.
- Mesleh [25] implements a Support Vector Machine (SVM) based text classification system for Arabic language articles. He used a corpus of online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram, and Al-Dostor. The system shows a high classification performance on the data set in terms of F-measure (F=88.11).
- Harrag *et al.* [19] presents the results of classifying Arabic text documents using a Decision Tree algorithm (DT). The study concluded that the effectiveness of the improved classifier is very good and gives generalization accuracy about 0.93 for the scientific corpus and 0.91 for the literary corpus.
- Artificial Neural Network (ANN) for the classification is also used, by Harrag *et al.* [19], to classify Arabic documents. For the used corpus, the performance achieves 88.75%.
- A comparative study of various classifiers (ME, NB, DT, ANN, SVM and KNN) has been made by El-Halees [13], using the same data set. He found that Naïve Bayes show the best F-measure accuracy (F1= 91.81), while Maximum Entropy, Support Vector Machine, and Decision Tree achieve an acceptable performance.

Table 1. Comparison of Classification algorithms.

Reference	Used classifier	Accuracy or F-measure
El Kourdi [14]	Naive Bayes	68.78 %
El-Halees [13]	Maximum entropy	80.41%
Sawaf [26]	Maximum entropy	62.7%
Mesleh [25]	Support Vector Machines	F=88.11
Harrag [19]	Decision tree	93%
Harrag [19]	Artificial Neural Network	88.33%
El-Halees [13]	Naïve Bayes	F=91.81

In all previous systems (Table 1), each author uses his own data set, for this reason we cannot make a decision about the best Arabic questions classifier. Regarding the non availability of the Arabic resources, each author uses his own dataset for testing his method. Therefore, we cannot make a true comparative study for the existing Arabic QASs. For this reason, most of the research on QA has been applied to English language. There are, however, interesting examples in other languages, including Arabic. In the following we present some Arabic QAS:

- QARAB, built in 2002 by [18] is a rule-based Factoid QA system for Arabic. This system works with unstructured data from documents collected from Al-Raya Newspaper with 113 Factoid questions. But, it didn't handle the two types of questions "How and Why" (كيف،لماذا).
- DefArabicQA [30] is an Arabic Definition Question Answering system which was introduced in 2010. This system answers questions of the form "What is ..?". This system uses the web as data source. To evaluate the system, two experiments were conducted with Google only and Google coupled with Wikipedia as the web sources.
- QArabPro [5] was developed in 2011. This system assumes that the answer must exist within one of the documents that were used as a corpus. However, this system does not handle " كيف" (How) type of questions.

3. Question Classification

Before classifying Arabic questions, the pre-processing phase is necessary. We remove all the punctuations, diacritics and stopwords, and then we tokenize questions. For doing this treatment, we use Alkhalil Morph syntactic Analysis System for Arabic Text (Alkhalil Morpho Sys) [10], which is one of the best open source morphological analyzers.

In this work, we adopt a machine learning approach to classify questions. We made a comparative study of the three most popular sentences classifiers:

- *A Support Vector Machine (SVM)*: is a machine learning algorithm that is based on statistical learning theory. SVM are linear functions of the form $y_i\ f(x) = < w \bullet x > +b$, where $< w \bullet x >$ is the inner product between the weight vector w and the input vector x. The SVM can be used as a classifier by setting the class to 1 if $f(x) > 0$ and to -1 otherwise. The main idea of SVM is to select a hyperplane that separates the positive and negative examples while maximizing the minimum margin, where the margin for an x_i is $y_i\ f(x)$ and $y_i \in$ [-1,1] is the target output. This corresponds to minimizing $<w \bullet w>$ subject to $y_i <w \bullet w> + b \geq$ for all i. Large margin classifiers are known to have good generalization properties. An adaptation of the A Library for Support Vector Machines (LIBSVM) implementation is used in the following. Four types of kernel function linear, polynomial, radial basis function, and sigmoid are provided by LIBSVM [11].
- *A Decision Tree (DT)*: is a tree whose internal nodes are tested and whose leaf nodes are categories. Each internal node test one attribute and each branch from a node selects one value for the attribute. The attribute used to make the decision is not defined. So we can use the attribute which gives maximum information. And the leaf node predicts a category or a class. Decision trees are not limited to boolean functions, but they can be extended to general categorically value functions. In Figure 2 the given instances can be divided based on the values it takes for the attribute "outlook". The instances are split based on attributes and the one which gives the maximum information is selected as the decision for that node. Hence, in the above example, we could say that selecting "Outlook" at the root node gives maximum information at that level. And the edges represent the values the attributes can take and the instances are divided accordingly to each child node. The tree can be many trees depending upon the values that the attributes can take. The attribute selection is based on a heuristic approach that the particular attribute will give the best split at a particular level. This approach has been successful over the past [28].

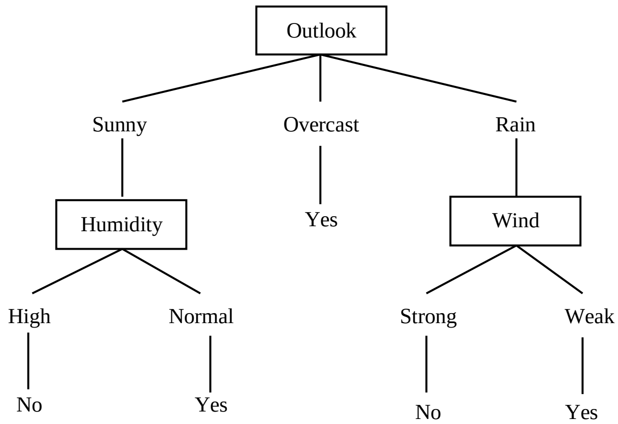

Figure 2. An example of a decision tree.

- In the Naive Bayes (NB) classifier each document is viewed as a collection of words and the order of words is considered irrelevant. The probability of a class c given a test document d is computed as follows:

$$P(c|d) = \frac{P(c)\prod_{\omega \epsilon d} P(\omega|c)^{n_{\omega d}}}{P(d)} \tag{1}$$

Where $n_{\omega d}$is the number of times word w occurs in document d, $P(c)$ is the probability of observing word w given class c, $P(c)$ is the prior probability of class c, and $P(d)$ is a constant that makes the probabilities for the different classes sum to one. $P(d)$ is estimated by the proportion of training documents belonging to the class c and $P(c)$ is estimated as follows:

$$P(\omega|c) = \frac{1 + \sum_{d \epsilon D_c} n_{\omega d}}{k + \sum_{\omega'} \sum_{d \epsilon D_c} n_{\omega' d}} \tag{2}$$

Where D_c is the set of all training documents of class c, and k is the size of the vocabulary (i.e., the number of distinct words in all training documents). The additional one in the numerator which is the so-called Laplace correction and corresponds to initializing each word count to one instead of zero. It requires the addition of k in the denominator to obtain a probability distribution that sums to one. This kind of correction is necessary because of the zero-frequency problem: a single word in the test document d that does not occur in any training document pertaining to a particular category c will otherwise render $P(c|d)$ zero [16].

4. From a Question to a Query

4.1. Taxonomy

Before classifying questions we adopt taxonomy. Taxonomy is an information classification method in a structured architecture.

In [23] Authors have been proposed four types of taxonomy, which cover all existing taxonomies:

- Taxonomies based on the type of interrogative question.
- Taxonomies based on the description style of the question.
- Taxonomies based on the semantic interpretation of the answer type.
- Taxonomies based on restricted domains.

The first one is based on the common type of interrogative questions, for example, seven coarse classes from English interrogative tools (ITs) have been proposed (who, where, what, when, which, why, how). In Arabic, linguists have defined 13 interrogative tools ITs (Table 2).

Table 2. Arabic Taxonomy.

Interrogative Tools	**Use (استعمال)**
Who (من)	Human (العاقل)
How (كيف، أنى)	Description (حال الشيء و هيئته)
Where (أين، أنى)	Location (المكان)
When (متى, أيان)	Time (الزمان)
How, many (كم)	Number (العدد)
What (أي، ما)	All above uses (يستفهم بها عن جميع ما تقدم)

They are divided into two sets:

1. Nouns “أسماء” (من، ما، أي، كم، كيف، متى، أيان، أين، أنى)
2. Particles “حروف” (هل, أ, أم)

The second taxonomy type is based on an interrogation style, for instance: Definition: “what does mean?”, Example: “what is an example label or instance of the category?”, Quantification: “how many?”.

In the taxonomies based on the semantic interpretation of the answer type, the semantic interpretation can be made on several levels. Li and Roth [24] have proposed taxonomy with a double level (Table 3).

Table 3. Li and Roth [24] taxonomy.

Coarse class	**Fine class**
Abbrev	Abbreviation, Expression abbreviated
Entity	Animal, Body, Color, Creative, Currency, Disease, Event, Food, Instrument, Language, Letter, Other, Plant, Product, Religion, Sport, Substance, Symbol, Technique, Term, Vehicle, Word.
Description	Definition, Description, Manner, Reason.
Human	Group, Individual, Title, Description
Location	City, Country, Mountain, Other, State
Numeric	Code, Count, Date, Distance, Money, Order, Other, Period, Speed, Temperature, Size, Weight.

Finally, the taxonomy for restricted domains, which depends on the treated domain. The used domain for this taxonomy type is the medical domain (Anatomy, diagnosis, cause disease, etc.,).

We are going to use the taxonomy based on the semantic interpretation of the answer type, since it is the most used taxonomy type used in question answering Text Retrieval Conference (TREC) conferences (QA track of TREC). Specifically, we will opt for the taxonomy proposed by Li and Roth [24] for the following reasons:

- It has two levels.
- It is an open domain taxonomy.
- It contains an interesting number of classes that have a positive influence on a QAS performance.

We will use also Arabic taxonomy (Table 2), because it’s the most used in Arabic text classification.

4.2. Features Extraction

Document representation is the task of representing a given document in a suitable form for data mining system. There are several ways in which the conversion of documents from plain text to instances with a fixed number of attributes. In [15], the authors describe the most known features to extract:

- *Bag-of-Words (BOW)*: is the most commonly used word-based representation method. We adopt this representation method in this work. With this representation a document is considered to be simply a collection of words which occur in it at least once. Using this approach, it is possible to have tens of thousands of words occurring in a fairly small set of documents. Many of them are not important for the learning task and their usage can substantially degrade performance. It is imperative to reduce the size of the feature space. One widely used approach is to use a list of common words that are likely to be useless for classification, known as stopwords, and remove all occurrences of these words before creating the BOW representation. Another very important way to reduce the number of words is to use stemming task which reduces inflected words to their stem and keeps the latter as a representational feature.
- *N-gram:* Word n-gram contextual features can be derived from the context of a document in order to extract the relationships between previously identified named entities (NEs), also used by [8] and [27], and an encountered word within the input document [23]. They are used to investigate the space of the surrounding context for the NEs by taking into account the features of a window of words surrounding a candidate word in the recognition process. Moreover, the character n-gram models attempt to capture surface clues that would indicate the presence or absence of an NE. For example, character bigram, trigram, and 4-gram models can be used to capture the prefix attachment of a noun for a candidate NE such as the determiner, a coordinating conjunction, a preposition, etc. On the other hand, these features can also be used to conclude that a word may not be an NE if the word is a verb that starts with any character of the verb present tense character.

Despite the fact that lexical features have solved the problem of dealing with a large number of prefixes and suffixes, they do not resolve the compatibility problem between prefixes, suffixes, and stems.

- *Part-of-Speech (POS)*: POS morpho-syntactic tag, which plays a significant role in Arabic NLP. An Arabic NE usually consists of either Common Noun (NN) or Proper Noun (NNP) tags. In [18] very good results were obtained using the POS tagging feature, which was exploited to improve NE boundary detection. The shared task of CoNLL now includes a POS column in its corpora. Thus, the POS tag is a good distinguishing feature for Arabic NEs.

4.3. Rewording Query

At this level, our question is transformed to a bag of words, without stop words, diacritics and punctuations. As mentioned in [22] paper, the meaning of a sentence is understood by the set of nouns contained in this sentence (not the verbs). For that reason, we use an Arabic POS Tagger to distinguish between nouns and verbs in a sentence.

POS tagging is a crucial step for several applications in natural language processing. It consists of attributing to each word of a given sentence a tag that provides useful information (type, gender, the tense for verbs, etc.,).

In the field of Arabic POS tagging, many studies have been made, we mention:

- *Arabic POS tagger library*: It was developed by Qatar Computing Research Institute (QCRI). It is based on the CRF++ model (conditional random fields) [12].
- *StanfordPOS*: This POS tagger was developed in 2003 for English [29]. It was later extended to other languages (Arabic, Chinese, German, French,…). It is constantly improved and freely distributed on the Stanford University website.
- *The hybrid Arabic POS tagger*: Proposed by [1], supports enclitic and proclitic attached to a given word. This system determines the syntactic function of proclitics. It combines the rule-based and statistical methods to allocate the correct tags for a given sequence of words.

The following table (Table 4) presents a comparison between those three systems, using the global metrics: the precision, the recall, the F-measure and the accuracy.

Table 4. Comparison between three Arabic POS tagging.

	Stanford POS %	**QCRI %**	**Hybrid Arabic POS tag %**
Precision	75.65	73.11	87.30
Recall	54.66	64.62	88.98
F-measure	63.46	68.60	88.13
Accuraccy	72.68	84.54	94.02

4.4. Query Expansion

Query Expansion can be defined as the process of reformulating the query to overcome the problem of mismatching relevant documents and to improve the performance of a search engine by retrieving more relevant documents [6].

Our query is represented by a set of nouns. The main idea of this section is to extend a query by the synonyms of each noun in the query, using Arabic WordNet (AWN).

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus.

AWN is a WordNet for Arabic language, it’s created in 2006 by [9] and it has been extended in 2013 by [3]. AWN is a lexical database of the Arabic language following the development process of Princeton English WordNet and Euro WordNet. It utilizes the Suggested Upper Merged Ontology as an interlingua to link Arabic WordNet to previously developed wordnets. AWN contains about 11,000 synsets (including 1,000 NE) [20].

In [2], authors have used AWN to extend Arabic query for an information retrieval system and the results were very encouraging. Table 5 presents an example of a query expansion using AWN:

Table 5. Arabic AWN expansion example.

Term	**Term Expansion**
مخطط	إستراتيجية، تصميم، مخطط ,ترتيب
تجهيز	تجهيز، توفير، تزويد
إبداع	إبداع، إبتكار ، تكوين، خلق
قانون	قانون، تشريع ، إجراء

Currently, it’s the most important Arabic lexical database and we use it in our system.

5. Search Engine

Information Retrieval (IR) is the process of obtaining relevant documents corresponding to a user query from a collection of information resources. The used tool to get the relevant documents called a search engine.

Google Web Search or simply Google is a web search engine developed by Google. It is the most-used search engine on the World Wide Web, handling more than three billion searches each day. As of February 2016, it is the most used search engine in the US with 64.0% market share [4].

Wikipedia is a multilingual, web-based, free-content encyclopaedia project supported by the

Wikimedia Foundation and based on a model of openly editable content. Wikipedia is the largest and most popular general reference work on the Internet and is named as one of the most popular websites. The project is owned by the Wikimedia Foundation, a non-profit organization which "operates on whatever money it receives from its annual fund drives". In this paper, Wikipedia represents our collection.

Let's remind the whole phases of our system before obtaining the relevant documents by Google search engine (Figure 3).

The following pseudo code describes the whole proposed algorithm:

Algorithm 1: Documents retrieval algorithm

/ Before any treatment we remove diacritics, punctuations and foreign characters (numbers, Latin characters) we keep only the Arabic ones. */*

```
{
Input: Question (a set of terms)
Output: Documents (a set of documents)
Question={ term 1, ..., term N} /* list of question's terms*/
SWL={sw 1, sw 2,..., sw K } /* list of stopwords */
Query={} /* initialization of a query's terms list */
C={} /*  list of the concepts of a question's term */
SVM: Support Vector Machine Classifier
POS: Arabic POS tagging
AWN: AWN Ontology
Begin
For each term ∈ Question do
{      if term ∈ SWL then omit term
       else add term to Query
}
Class=getClass_SVM(Query) /* Use SVM classifier to classify Query */
add class's name to Query
/* add the class's name to Query as a new term */
For each term ∈ Query do
{      Use POS for term
        tag=get.tag(term)
       if (tag = verb) then delete term from Query
 }
L= Query.getLength
For i=1 to L
{     Map the term i in AWN
      C= getConcepts(i)
      if C is not empty then add the concepts C to Query
}
/* the query is extended by the set of the concepts of each term of the question */
Call GoogleAPI with Arabic Wikipedia
Documents=search(Query)
End
}
```

Our system contains two phases:

- Question processing.
- Information retrieval.

In the first phase, we remove diacritics, punctuations and foreign characters (numbers, Latin characters) we keep only the Arabic ones. We remove also stop words. For the learning step we use SVM classifier to get the class of a new question. The name of the detected class is added to the question terms to formulate a new query.

In the second phase, we remove verbs from a query using hybrid Arabic POS tagging [1]. We expand the query using AWN, by retrieving a concepts' set of each term. We integrate Google as a search engine and Arabic Wikipedia as a dataset.

6. Experiments

In order to test the three classification algorithms, we use a fusion of two sets: TREC and CLEF. Because of the lack of Arabic resources, we use the Arabic translation of TREC and CLEF datasets. Because of this translation, we had some Language issues such us abbreviation, which doesn't exist in Arabic, but we use it as long as we don't have a special dataset. We annotate manually the dataset (Table 6).

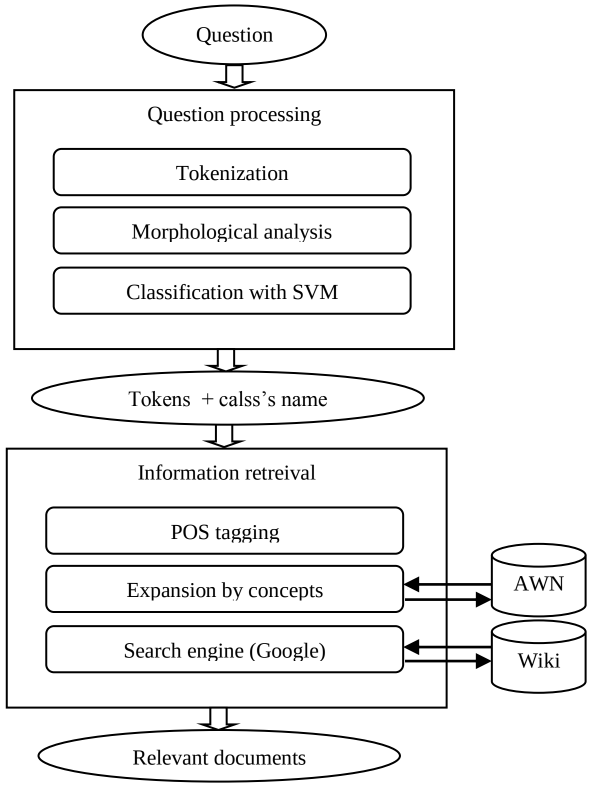

Figure 3. Flowchart of our proposed method.

Table 6. Annotated Classes.

Question type		Number
Abbreviation		24
Definition		150
Description		80
Location	**City**	80
	Country	60
	Other location	275
Person		420
Time		300
Number		270
Entity		230
Other		591
Total		2300

The dataset contains questions in different domains with a different ITs. But it does not cover all domains expressed by our taxonomies, especially for the fine classes in Li and Roth [24] taxonomy. We will test the performance using only the types: abbreviation, definition, description, city, country, other location, person, time, number, and entity.

We test the three algorithms (SVM, Decision Tree and Naive Bayes) using a bag-of-words as a feature of extraction.

Our data are a set of non annotated questions, we attribute manually the expected classes for 60% of questions and we consider it as a training data. The other 40% constitute our test data.

Table 7. Examples of questions of the used data set.

Question	Class (manually annotate)	Class detected with		
		SVM	Naïve Bayesian	Decision-tree
من كان أول شخص وصل الى القطب الجنوبي؟ Who was the first person to reach the South Pole?	Person	Person	Non detected	Person
من هي ملكة المملكة المتحدة ؟ Who is the Queen of the United Kingdom?	Person	Person	Person	Person
كم يبلغ ثمن تذكرة التيتانيك ؟ How much does the Titanic cost?	Number	Number	Number	Non detected
ما هو قطر كرة الغولف ؟ What is golf ball diameter?	Number	Number	Definition	Definition
ما هي درجة انصهار النحاس ؟ What is the degree of copper fusion?	Number	Number	Number	Non detected
ما هي عاصمة ولاية ويسكونسن؟ What is the capital of Wisconsin?	City	City	City	City
ما هي أكبر مدينة في الولايات المتحدة الأمريكية ؟ What is the largest city in the United States of America?	City	City	City	Person
ما هو مرض الأوتيزم ؟ What is Autism?	Disease	Non detected	Description	Definition
ما هي الألوان الأولية التي يجب مزجها للحصول على اللون البرتقالي ؟ What are the primary colors that must be mixed to get orange?	Color	Non detected	Definition	Non detected
ما هو اف دي أي ؟ What is FDI?	Abbreviation	Abbreviation	Definition	Non detected

Table 7 shows examples of the question class detection. We have three cases for a question class detection:

- The question's class is correctly detected.
- The question's class detected is incorrect.
- The question's class is not detected.

Table 8. The average performance of the tested classifiers.

	Taxonomy	Time efficiency (ms)
SVM	Arabic taxonomy	45
	Li and Roth [24] Taxonomy	65
NB	Arabic taxonomy	40
	Li and Roth [24] Taxonomy	55
DT	Arabic taxonomy	60
	Li and Roth [24] Taxonomy	70

In Table 8 we compare the time efficiency of the three classifiers (SVM, NB and DT). The given results represent the average time of each classifier on the test phase (not the training test). The results are calculated with millisecond (ms). The time efficiency of SVM and NB is almost the same but DT time is a little longer.

In order to make a decision about the chosen algorithm, we make statistical study using recall, precision and F1-measure. The following table (Table 9) presents the obtaining results.

Table 9. Comparison of the obtained results in terms of recall, precision and F1-measure using SVM, NB and DT classifiers.

	Taxonomy	Recall	Precision	F1-Measure
SVM	Arabic Taxonomy	0.89	0.93	0.90
	Li and Roth [24] taxonomy	0.82	0.85	0.83
NB	Arabic Taxonomy	0.80	0.79	0.79
	Li and Roth [24] taxonomy	0.77	0.80	0.78
DT	Arabic taxonomy	0.80	0.84	0.81
	Li and Roth [24] taxonomy	0.58	0.78	0.66

Concerning our data, we note clearly that SVM classifier gives the best results for our Arabic data set with F1-measure=0.90 for Arabic taxonomy and 0.83 for Li and Roth [24] taxonomy.

Because of the lack of Arabic resources, we cannot make a precise comparative study of machine learning questions classification; because each author uses his own dataset to classify the Arabic questions.

After the classification step, we use a POS tagging process to remove verbs and AWN to expand query by concepts. To extract an answer we use a search engine to get relevant documents from the dataset.

The Figure 4 describes a real example from our dataset.

Our proposed methodology seems to work well because of many reasons. We use SVM classifier which achieves the better performance (than DT and NB) for our dataset; the detected class is added to the input query as a new features. For a question reformulation, a good POS tagger is used, based on [1] study on 2015, the hybrid Arabic POS tagging gives the best results, comparing with Arabic POS Tagger Library and StanfordPOS. We extended query term by concepts from AWN. We use a powerful search engine ‘Google’ and our dataset is Wikipedia. Translation techniques could produce many problems, for that reason, we use the Arabic version of Wikipedia.

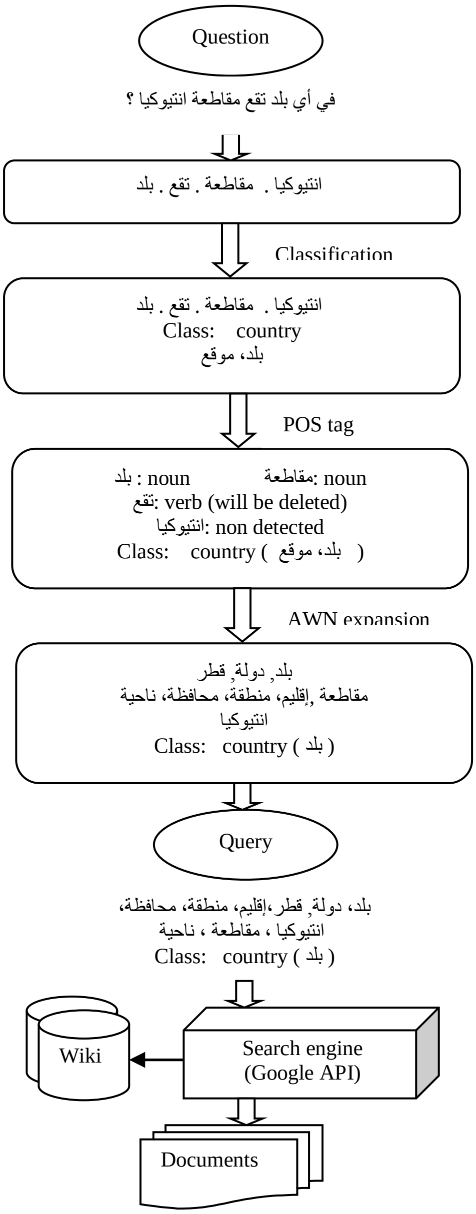

Figure 4. Eexample of obtained result for a question.

7. Conclusions and Future Work

The great importance of a QAS is that it returns one precise answer to a user question, contrary to search engine. This question is written in natural language, and this makes it complicated for processing. NLP regroups a lot of techniques to extract meaning from a human speech. Arabic NLP still not meets all the needs unlike languages that use Latin script. Therefore, many recent researchers use machine translation techniques [7]. Using machine translation, we can resolve many problems, but we can also lose the true meaning of the sentence written in natural language.

For classifying Arabic questions, we present, in this paper, a comparative study of three different machine learning classifiers (DT, SVM, NB). For our data set, the experimental results show the efficiency of the SVM classifier with 84 % as a percentage of the correct class detection.

To transform the natural question to a query, we used a hybrid Arabic POS tagging for annotating each token in the query. Then, we used Arabic WordNet to expand the query nouns by synonyms. The strong point of this research is that we don’t use translation to another language in all system phases. To get an answer, we extract first relevant documents using Google API and Arabic Wikipedia as a collection.

As a future work, we are going to use disambiguation strategies to get the right concepts of the query terms. The terms are mapped to their corresponding concepts using different strategies, which are described on [17], for adding terms by concepts. Then we will extract passages from the relevant documents.

References

[1] Ababou N. and Mazroui A., “A Hybrid Arabic POS Tagging for Simple and Compound Morphosyntactic Tags,” *International Journal of Speech Technology*, vol. 19, no. 2, pp. 289-302, 2016.

[2] Abderrahim M., Abderrahim M., and Chikh M., “Using Arabic Wordnet for Query Expansion in Information Retrieval System,” *in Proceedings of the 3rd International Conference on Web and Information Technologies IEEE,* Marrakech, 2010.

[3] Abouenour L., Bouzoubaa K., and Rosso P., “On the Evaluation and Improvement of Arabic Wordnet Coverage and Usability,” *Language Resources and Evaluation Journal*, vol. 47, no. 3, pp 891-917, 2013.

[4] Adam L., “Desktop Search Engine Rankings,” Technical Report, 2016.

[5] Akour M., Abu Fardeh S., Magel K., and Al-Radaideh Q., “QArabPro: A Rule Based Question Answering System for Reading Comprehension Tests in Arabic,” *The American Journal of Applied Sciences*, vol. 8, no. 6, pp. 652-661, 2011.

[6] Al-Shalabi R., Kanaan G., Yaseen M., Al-Sarayreh B., and Al-Naji N., “Arabic Query Expansion using Interactive Word Sense Disambiguation,” *in Proceedings of the 2nd International Conference in Arabic Language Resources and Tools*, Cairo, 2009.

[7] Bakari W., Bellot P., and Neji M., “A Logical Representation of Arabic Questions Toward Automatic Passage Extraction from the Web,” *International Journal of Speech Technology*, vol. 20, no. 2, pp. 339-353, 2017.

[8] Benajiba Y., Zitouni I, Diab M., and Rosso P., “Arabic Named Entity Recognition: Using Features Extracted from Noisy Data,” *in Proceedings of the Association for*

Computational Linguistics Conference Short Papers, Uppsala, pp 281-285, 2010.

[9] Black W., Elkateb S., Rodriguez H., Alkhalifa M., Vossen P., Pease A., Bertran M., and Fellbaum C., "The Arabic WordNet Project," *in Proceedings of the International Conference on Language Resources and Evaluation*, Genoa, 2006.

[10] Boudchiche M., Mazroui M., Bebah M., and Lakhouaja A., "L'analyseur Morphosyntaxique Alkhalil Morpho Sys 2," *1ère Journée Doctorale Nationale sur L'Ingénierie de la Langue Arabe*, Rabat, 2014.

[11] Chang C., and Lin C., "A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology Journal*, vol. 2, no. 3, pp. 1-27, 2011.

[12] Darwish K., Abdelali A., and Mubarak H., "Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging," *in Proceedings of the International Conference on Language Resources and Evaluation*, Iceland, 2014.

[13] El-Halees A., "Arabic Text Classification using Maximum Entropy," *the Islamic University Journal-Series of Natural Studies and Engineering*, vol. 15, no. 1, pp. 157-167, 2007.

[14] El-Kourdi M., Bensaid A. and Rachidi T., "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, pp. 51-58, 2004.

[15] Elsebai A., Meziane F., and Belkredim F., "A Rule Based Persons Names Arabic Extraction System," *in Proceedings of the 11th International Business Information Management Association Conference*, Cairo, 2009

[16] Frank E. and Bouckaert R., "Naive Bayes for Text Classification with Unbalanced Classes," *in Proceedings of the 10th European conference of Knowledge Discovery* in Databases, pp 503-510, Berlin, 2006.

[17] Hadni M., El Alaoui S., and Lachkar A., "Word Sense Disambiguation for Arabic Text Categorization," *The International Arab Journal of Information Technology*, vol. 13, no. 1A, pp. 215-222, 2016.

[18] Hammo B., Abu-Salem H., and Lytinen S., "QARAB: A Question Answering System to Support the Arabic Language," *in Proceedings of the workshop on Computational Approaches to Semitic Languages*, Philadelphia, pp. 1-11, 2002.

[19] Harrag F., El-Qawasmeh E., and Pichappan P., "Improving Arabic Text Categorization Using Decision Trees," *in Proceedings of the First International Conference on Networked Digital Technologies*, Ostrava, pp. 110-115, 2009.

[20] Weisscher A., http://www.talp.upc.edu/content/awn-arabicwn, Last Visited, 2014.

[21] Krutchen P., *The Rational Unified Process Model: An Introduction*, Addison-Wesley Professional, 2004.

[22] Lahbari I., Ouatik S. E., Zidani K. A., "A Rule-based Method for Arabic Question Classification," *in Proceedings of The International Conference on Wireless Networks and Mobile Communications*, Rabat, pp. 1-6, 2017.

[23] Li X., Xuan-Jing H., and Li-De W., "Question Classification using Multiple Classifiers," *in Proceedings of the 5th Workshop on Asian Language Resources (ALR-05) and First Symposium on Asian Language Resources Network*, Jeju Island, 2005.

[24] Li X. and Roth D., "Learning Question Classifiers: The Role of Semantic Information," *Natural Language Engineering*, vol. 12, no. 3, pp. 229-249, 2002.

[25] Mesleh A., "Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study," *in Proceedings of the 12th WSEAS International Conference on Applied Mathematics*, Cairo, pp. 228-233, 2007.

[26] Sawaf H., Zaplo J., and Ney H., "Statistical Classification Methods for Arabic News Articles," *in Proceedings of the Workshop on Arabic Natural Language Processing*, Toulouse, 2001.

[27] Shaalan K., "A Survey of Arabic Named Entity Recognition and Classification," *Computational Linguistics*, vol. 40, no. 2, pp. 469-510, 2014.

[28] Srinivasan R., "Multiclass Text Classification A Decision Tree based SVM Approach," Technical Report, 2006.

[29] Toutanova K., Klein D., Manning C., and Singer Y., "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," *in Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, pp. 173-180, 2003.

[30] Trigui O., Belguith L., and Rosso P., "DefArabicQA: Arabic Definition Question Answering System," *in Proceedings of the 7th Workshop on Language Resources and Human Language Technologies for Semitic Languages*, Valletta, pp. 40-45, 2010.

Imane Lahbari is a Phd student in Laboratory of Informatics and Modeling, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco. She received the master degree in Information Systems, Network and Multimedia in 2015. Her research interests include Natural language processing and Question Answering Systems.

Said El Alaoui is working as a Professor since 1997 in Department of Computer Science, Faculty of Sciences Dhar EL-Mahraz (FSDM) at Sidi Mohamed Ben Abdellah University (USMBA), Fez, Morocco. His current research interests include Natural Language Processing, Information Retrieval, Biomedical Question Answering, Biomedical Information Extraction, and Arabic Document Clustering and Categorization, High-dimensional indexing and Content-Based Image Retrieval.

Khalid Zidani received his PhD degree from the Faculty of Sciences, Nantes, France, in 1994 in computer Science. His current research interests include Natural Language Processing, Arabic Text Mining, Information Retrieval, Document Clustering and Categorization, Content-Based Image Retrieval, Large Image Databases Indexing.