ACIT'2017
The International Arab Conference on Information Technology
Yassmine Hammamet, Tunisia
December 22-24, 2017

Arabic Question Classification using Machine Learning Approaches

Imane Lahbari, Said El Alaoui Ouatik and Khalid Alaoui Zidani
Sidi Mohamed Ben Abdellah University, Morocco

***Abstract*—Question Classification is a very important component in question answering system. In this paper, we present a machine learning comparison study for classifying Arabic questions. We have used two taxonomies: Arabic taxonomy and Li & Roth taxonomy. We have conducted several experiments using TREC and CLEF questions.**

Keywords— Natural Language processing; Arabic Question Answering System; Question Classification; Taxonomy; Machine learning Approach; SVM; Decision-tree; Naive Baye

I. 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, these 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. Thus, seeking 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 largest 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 words. In this regard, this language requires specific 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. Colloquial Arabic depends, to some extent, upon the dialects spoken in each region. The Arabic dialects are spoken in informal settings.

The next figure presents the flow chart of QAS with 3 blocs:

- Question preprocessing
- Information retrieval
- Answer Processing

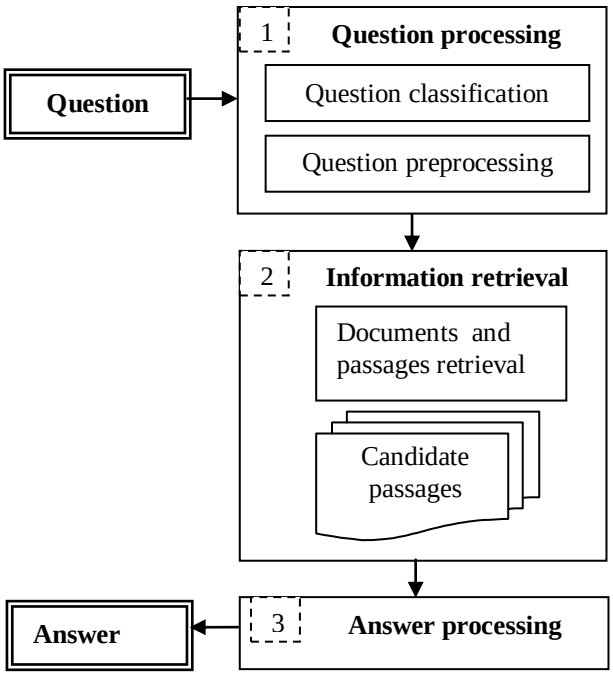

Fig. 1. QAS architecture

The Arabic Question classification plays a vital role in QAS, it 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 [1] taxonomy.

This paper is organized as follows: section II presents the related work. In section III we describe our proposed study

about classifying questions using machine learning approaches. Experiments are discussed in section IV and we conclude in section V with discussing the future works.

II. Related Work

A 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 a specific rules depending patterns. The second, machine learning approach, classify questions after a learning step which need an annotated data set.

In this work we have adopted the machine learning approach, because it:

- Is the most used
- Cover all question types
- Flexible for the new data
- Less complicated than 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 Koudri [1] classified Arabic web documents automatically by Naive Bayes (NB) which is a machine learning algorithm. Cross validation experiments were used to evaluate the NB 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 [2] and Sawaf [3] for classifying Arabic news articles. The classification accuracy was 80.41% and 62.7% by Sawaf without any morphological analysis.
- Meshleh [4] implements a Support Vector Machine (SVM) based text classification system for Arabic language articles. He used an in a corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram, and Al-Dostor. The system shows a high classification effective for Arabic data set in term of F-measure (F=88.11).
- Harrag in [5] 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 a 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 in [5], to classify Arabic documents. For the used corpus the performance attends 88.75%.
- Halees in [6] mad a comparative study for six classifiers: ME, NB, DT, ANN, SVM and KNN, with the same data set. He found that the performance of Naïve Bayes is the best (F1= 91.81), the performance of Maximum Entropy, Support Vector Machine and Decision Tree are acceptable.

TABLE I. Comparison Of Classification Algorithms

Reference	Used classifier	Accuracy or F-measure
El Koudri [1]	Naive Bayes	68.78 %
El-Halees [2]	Maximum entropy	80.41%
Sawaf [3]	Maximum entropy	62.7%
Meshleh [4]	Support Vector Machines	F=88.11
Harrag [5]	Decision tree	93%
Harrag [5]	Artificial Neural Network	88.33%
Halees [6]	Naïve Bayes	F=91.81

In all previous systems, each author uses his own dataset, for this rason 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. For that reason we cannot make a comparative study for the existing Arabic QASs. I propose building one data set, available for everyone, which cover all types of question in all categories. At this time we can measure the evolution of Arabic QASs.

III. Our Study

A. *Taxonomy*

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

In [9] authors have 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, in [10] they proposed seven coarse classes from English interrogative tools (ITs) (who, where, what, when, which, why, how). In Arabic, linguists have defined 13 interrogative tools ITs (Table 2.). They are divided into two sets:

- Nouns "أسماء" (من، ما، أي، كم، كيف، متى، أيان، أين، أنىّ)
- Particles "حروف" (أم، أ، هل)

TABLE II. ARABIC TAXONOMY

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

The second taxonomy type is based on an interrogation style, in [11] they used this type of taxonomy, they have proposed 18 classes (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 [1] have been proposed taxonomy with a double level (Table 3.).

TABLE III. LI & ROTH [1] 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. It was used by [12] with 14 principal classes for the medical domain (Anatomy, diagnosis, cause...).

We are going to use the taxonomy based on the semantic interpretation of the answer type, since it is the most used taxonomy types used in question answering TREC conferences (QA track of TREC). Specifically, we will opt for the taxonomy proposed by Li and Roth [1] for the following reasons:

- Taxonomy in two levels.
- Open domain taxonomy
- Contains an interesting number of classes that have a positive influence on a QAS performance.

We will use also Arabic taxonomy, because it's the most used in Arabic text classification.

B. *Feature extraction*

Document representation is the task of representing a given document in a form which is suitable 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 [9], the authors describes the most known features extracted to classify questions :

- Bag-Of-Words (BOW) is the most commonly used word-based representation method. With this representation a document is considered to be simply a collection of words which occur in it at least once. With 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 BOW representation. Another very important way to reduce the number of words is to use stemming which removes words with the same stem and keeps the stem as the 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 NEs and an encountered word within the input document [10]. 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,... 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.
- Base-Phrase Chunks (BPC): The structure of an Arabic sentence allows different arrangements of NEs:

NEs may appear anywhere in the sentence and at different distances from lexical triggers. [11] point out that these arrangements might complicate the structure of the induced heuristics. Rules of their rule-based NER system. This observation has led to using the BPC feature as an indicator of embedded NEs [10]. BPC features are related to the type of words that occur with NEs and their syntactic relations. They are usually identified by shallow syntactic parsing.

- Part-Of-Speech (POS): One of these features is the POS morpho-syntactic tag, which plays a significant role in Arabic NLP. An Arabic NE usually consists of either noun (NN) or proper noun (NNP) tags. In [12] 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.

C. Classification algorithms

SVM is a kind of machine learning approach based on statistic learning theory. SVM are linear functions of the form f (x) = <w*x> + b, where <w*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 example x i (i:index of x) is y i f(x) and y i ∈ [-1,1] is the target output. This corresponds to minimizing <w*w> subject to yi (<w*x> +b) >= for all i. Large margin classifiers are known to have good generalization properties. An adaptation of the LIBSVM implementation [a] is used in the following. Four types of kernel function linear, polynomial, radial basis function, and sigmoid are provided by LIBSVM. [13]

A **Decision Tree** 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 class. The decision trees are not limited to boolean functions, but they can be extended to general categorically value functions.

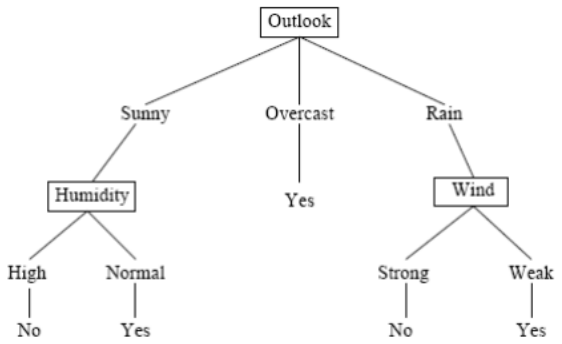

Fig. 2. An example of a decision tree

In the above example 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 a mary tree 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. But this approach has been successful over the past.[15]

In the **Naive Bayes** MNB classifier each document is viewed as a collection of words and the order of words is considered irrelevant. The probability of a class value c given a test document d is computed as

$$P(c|d) = \frac{P(c)\prod_{w\in d} P(w|c)^{n_{wd}}}{P(d)} \qquad (1)$$

where nwd is the number of times word w occurs in document d, P(w|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(c) is estimated by the proportion of training documents pertaining to class c and P(w|c) is estimated as

$$P(w|c) = \frac{1 + \sum_{d\in D_c} n_{wd}}{k + \sum_{w'} \sum_{d\in D_c} n_{w'd}} \qquad (2)$$

where Dc is the collection of all training documents in 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 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 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]

IV. Experiments

In order to test the three classification algorithms, we use a fusion of two sets: TREC (Text Retrieval Conference) [15] and CLEF (Cross Lingual Evaluation Forum) [16]. 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, abbreviation for example, don't exist in Arabic, but we use it, as long as we don't have a special dataset.

The dataset contains questions in different domains with a different ITs. But it's still don't cover all domains expressed by our taxonomies, especially for the fine classes in Li & Roth taxonomy. We will test the performance using only the types: Abbreviation, Definition, Description, City, Country, Other location, Person, Time, Number, and Entity.

For measuring the accuracy, we annotate manually the data (Table 4.).

TABLE IV. 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

We test the three algorithms (SVM, Decision Tree and Bayesian) using 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% will be our test data.

The next table (TABLE) presents some examples of questions detection classes:

TABLE V. AN EXAMPLE FROM 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	Person
ما هي أكبر مدينة في الولايات المتحدة الأمريكية ؟ What is the largest city in the United States of America?	Disease	Non detected	Description	Definition
ما هو مرض الأوتيزم ؟ What is Autism?	Color	Non detected	Definition	Non detected
ما هي الألوان الأولية التي يجب مزجها للحصول على اللون البرتقالي ؟ What are the primary colors that must be mixed to get orange?	Abbreviation	Abbreviation	Definition	Non detected
ما هو اف دي أي ؟ What is FDI?				

The preceding table shows us an example of the question class detection, we have three cases for a question class detection:

- The question class is correctly detected
- The question class detected is incorrect
- The question class is not detected

The next table presents the results:

TABLE VI. EXPERIMENTAL RESULTS

	Taxonomy	Percentage of non-detected and incorrect classes detection	Percentage of the correct Classes detection
SVM	Question Type	16%	84%
	Li & Roth	27%	73%
Naïve Bayesian	Question Type	24%	76%
	Li & Roth	35%	65%
Decision Tree	Question Type	37%	63%
	Li & Roth	50%	50%

About our data, we see clearly that SVM classifier gives the best results.

We can observe clearly that SVM classifier gives the best results for our Arabic data set (84 % as a percentage of the correct class detection).

Because of the lack of Arabic resources, we can't have the same data set and make a comparison between our results and the results of the related works using machine learning to classify questions.

V. Conclusion And Future Work

There are two main approaches for classifying questions: rule-based approach and machine learning approach. In this paper, we adopt a machine learning approach.

In this paper, we present a different classifiers, for our data set, the experimental results show the efficiency of the SVM classifier with 84 % as a percentage of the correct class detection.

This work presents a first step to building an Arabic question answering system, the next step is how to expand a query to get the most important documents and passages at the information retrieval phase.

References

[1] El-Kourdi M., Bensaid A. and Rachidi T. (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics. August, Geneva. Pp51—58

[2] El-Halees A. (2007), "Arabic Text Classification Using Maximum Entropy", The Islamic University Journal (Series of Natural Studies and Engineering), 15(1), pp. 157-167

[3] Sawaf H., Zaplo J., Ney H. (2001), "Statistical Classification Methods for Arabic News Articles", In the Workshop on Arabic Natural Language Processing, ACL'01, Toulouse, France.

[4] Mesleh A. (2007), "Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study", In the 12th WSEAS Int. Conf. on APPLIED ATHEMATICS, Cairo, Egypt.

[5] Harrag F., El-Qawasmeh E., Pichappan P. (2009), "Improving Arabic text categorization using decision trees", In the 1st Int. Conf. of NDT '09, pp. 110 – 115

[6] El- Halees A (2008)., "A Comparative Study on Arabic Text Classification", Egyptian. Computer Science Journal 20(2).

[7] a Jindal, Rajni, Malhotra, Ruchika, & Jain, Abha (2015). "Techniques for text classification: Literature review and current trends." Webology, 12(2), Article 139. Available at: http://www.webology.org/2015/v12n2/a139.pdf

[8] Khaled Shaalan , "A Survey of Arabic Named Entity Recognition and Classification", School of Informatics, University of Edinburgh, UK The British University in Dubai, UAE, Computational Linguistics Volume 40, Number 2

[9] R. Jindal, R. Malhotra, A. Jain, "Techniques for text classification: Literature review and current trends"

[10] Benajiba, Yassine, Mona Diab, and Paolo Rosso. 2008a. Arabic named entity recognition: An SVM-based approach.

[11] Elsebai, Ali, Farid Meziane, and Fatma Belkredim. 2009. A rule based persons names Arabic extraction system. In Proceedings of the 11th International Business Information Management Association Conference (IBIMA 2009)

[12] Benajiba, Yassine, Imed Zitouni, Mona Diab, and Paolo Rosso. 2010. Arabic named entity recognition: Using features extracted from noisy data

[13] C. C. Chang, C. J. Lin, "LIBSVM: A Library for Support Vector Machines"

[14] LI Xin, HUANG Xuan-Jing, WU Li-de, "Question Classification using Multiple Classifiers"

[15] Srinivasan Ramaswamy, "Multiclass Text Classification A Decision Tree based SVM Approach"

[16] Eibe Frank, Remco R., Bouckaert, "Naive Bayes for Text Classification with Unbalanced Classes"