ARTICLE IN PRESS

Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An arabic question classification method based on new taxonomy and continuous distributed representation of words

Alami Hamza, Noureddine En-Nahnahi, Khalid Alaoui Zidani, Said El Alaoui Ouatik

Sidi Mohammed Ben Abdellah University, Laboratory of Informatics and Modeling (LIM), Sciences Faculty Dhar El Mahraz, PO Box 1796, Fez 30003, Morocco

ARTICLE INFO

Article history:
Received 15 August 2018
Revised 25 December 2018
Accepted 4 January 2019
Available online xxxx

Keywords:
Arabic question answering
Arabic taxonomy
Question classification
Natural language processing
Question distributed representation
Word embeddings
Machine learning

ABSTRACT

The unability of search engines to retrieve precise answer for a given question leads research teams to build question answering systems (QAS). These systems provide exact answers of questions formulated in natural languages. Question classification is a crucial task for QAS since finding the correct answer type increases the performance of this latter. The questions taxonomy plays an important role in question classification. A broad range of taxonomies are proposed; most of these are not designed for Arabic questions. The contribution of the paper is twofold. First, we build a taxonomy for open domain Arabic questions. Second, we propose an efficient method for classifying Arabic questions. The basic idea consists of two stages: first, we compute representation of questions according to continuous distributed representation of words which allows to capture syntactic and semantic relations between words. Then, we apply a machine learning approach to classify questions into seven types or categories. We carried out several experiments and compared the proposed method with different state of arts Arabic question classification methods. Experimental results show that the proposed method achieves 90% in terms of accuracy.

1. Introduction

Nowadays an enormous amount of textual documents is continuously produced in different languages and requires advanced and efficient search facilities that satisfy the user needs. QAS are designed to facilitate the process of seeking information by automating the extraction of the desired information. The behavior of these systems should be close to the human behavior in order to meet what the user is looking up and to give the relevant answer. Indeed, the QAS input is a natural language question, e.g., *"When was Fez city built?"* and its output is a natural language answer, e.g., *"Fez was built in 789"*. Fig. 1 shows the three main modules of question answering system: (1) Question processing; (2) Passage retrieval; (3) Answer processing.

1. Question processing: This component performs two tasks: Keywords extraction and question classification. From the one hand, keywords are mandatory for the passage retrieval module and on the other hand, question type is essential for the answer processing module.
2. Passage retrieval: In this component, one can first perform an IR system for retrieving the relevant documents based on the keywords provided by the question processing module. Then, information extraction (IE) techniques are employed for extracting the candidate passages which contain the possible responses. These passages are fed into the answer processing module.
3. Answer Processing: In this module, the extracted passages as well as the question type are used together in order to provide the final answers formulated in natural language.

Intuitively, the question processing module plays an important role in QAS. To be sure, enhancing the performance of the keywords extraction and question classification tasks obviously affects positively the relevance of both the retrieved passages and the final answers. Moldovan et al. (2003) showed that 43.5% of the QAS failures are caused by the poor performance of the question processing module. This includes 36.4% of the failures related to question classification and 7.1% related to the keywords extraction task. These statistics illustrate the crucial role of the question classification task in QAS.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.01.001
1319-1578/

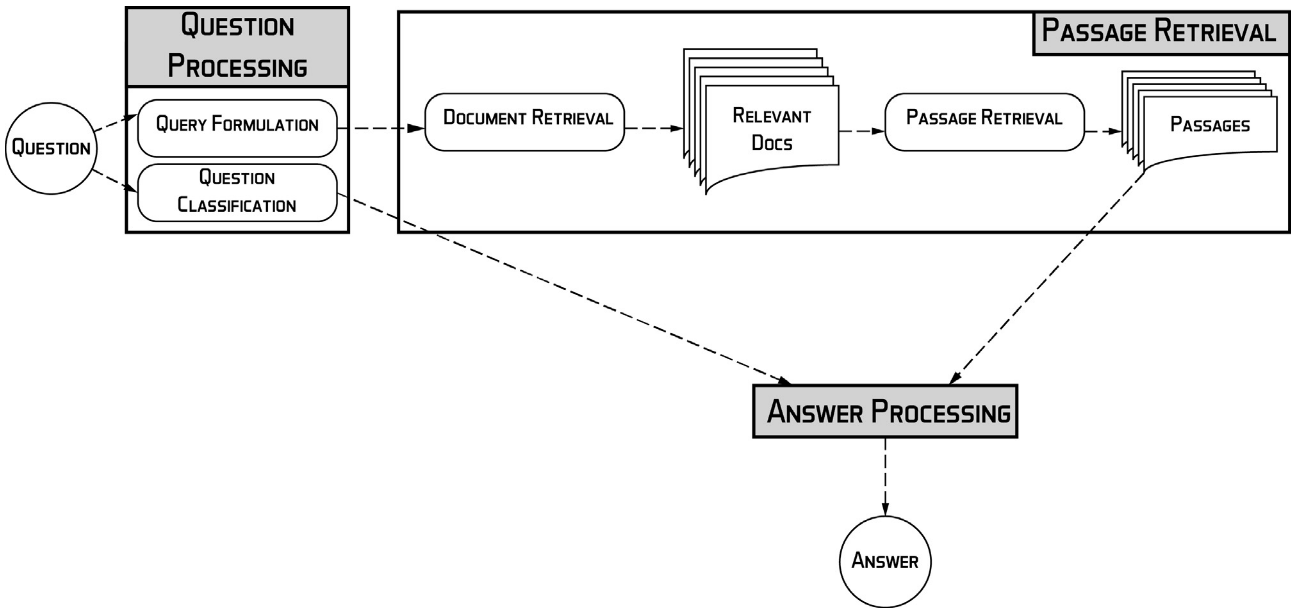

Fig. 1. Flowchart of question answering system showing three main modules: question processing, passage retrieval, and answer processing.

Numerous systems have been proposed in the literature for classifying both open-domain and restricted-domain questions in latin-based languages (Mishra and Jain, 2016; Li and Roth, 2002; Sarrouti and El Alaoui, 2017). Nevertheless, Arabic question classification faces many challenges due to the difficulties related to the complex morphology of this language, such as its derivational and inflectional nature, the presence of diacritical marks, the absence of capital letters and the lack of Arabic resources, e.g., corpora and taxonomies.

The aim of the paper lies on the proposition of a new taxonomy for Arabic question classification where other Arabic questions taxonomies are inspired from Wh questions or are for restricted domains. In addition, we explore new question representation (Bojanowski et al., 2016) in the domain of Arabic question classification which allows a reduced-size representation of words and captures syntactic and semantic relations between words.

Experiments are conducted to highlight the effectiveness of our method using a hand-crafted corpus from TREC,[1] CLEF[2] and Moroccan school books. For comparison purposes, we also use the latin-based Li & Roth taxonomy (Li and Roth, 2002), the TF-IDF representation method as well as the well-known classification algorithms, i.e., support vector machine (SVM), extreme gradient boosting (XGBoost), and others. The obtained results show the effectiveness of our Arabic question classification method.

The rest of the paper is organized as follows: Section 2 overviews the related work on Arabic question classification; Section 3 details the word representation adopted in this work; Section 4 presents our Arabic taxonomy and describes the proposed method for question classification; Section 5 presents the experimental results; Finally, Section 6 concludes and summarizes future prospects.

2. Related work

In this work, we focus on Arabic questions classification methods that enhance the performance of Arabic question answering systems (AQAS) which can be dated to 1993 (Mohammed et al., 1993). Generally, there are three approaches to classify questions: (1) Rule-based approach which is based on matching questions by implementing hand-crafted language rules; (2) Learning-based approach performs question classification by feeding questions features to a classifier and applying the trained classifier to predict the question type; (3) Hybrid approach combines the strengths of both rule-based and learning-based approaches.

First, we present a comprehensive survey of available rule-based Arabic question classification methods:

- QARAB (Hammo et al., 2002): One of the first AQAS to exploit Salton's (Salton, 1971) vector space model. The system used information retrieval and natural language processing techniques to get the best answer to a question. Their question classification method is based on the type of question particles. Questions are classified according to a set of known 'question types'.
- ArabiQA (Benajiba et al., 2007): The authors designed their own Arabic Named Entity Recognition (NER) system. This latter is used in several modules in their QAS including question classification.
- DefArabicQA (Trigui et al., 2010): A definitional question answering system which answer to questions in the form "*What is X?*". They considered question analysis module a vital component. The question type is defined from the interrogative pronoun of the question.
- IDRAAQ (Abouenour et al., 2012): The authors participated to question answering for machine reading evaluation (QA4ME) at CLEF 2012.[3] The question classification is based on interrogative particles.
- Al Chalabi et al. (2015) adopted a rule-based approach and used Arabic grammar rules to classify questions. The rules are constructed by the NOOJ tool.[4] They used 200 questions for training and 200 questions for test. They obtained 93% on recall and 100% on precision.

Most of rules presented in Abouenour et al. (2012), Hammo et al. (2002), Al Chalabi et al. (2015), Trigui et al. (2010) are based on interrogative particles, where each particle can identify the

[1] http://trec.nist.gov/.
[2] http://www.clef-initiative.eu/.
[3] http://clef2012.clef-initiative.eu/.
[4] http://www.nooj-association.org/.

Please cite this article as: A. Hamza, N. En-Nahnahi, K. A. Zidani et al., An arabic question classification method based on new taxonomy and continuous distributed representation of words, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.001

correct answer type. For instance, the interrogative particle "كم" (How) identifies the question of type number. However, the Arabic particles "ما، أي" (What) may yield to ambiguities. These two particles can be employed with different types of questions, such as definition, country, person, and so on. Unfortunately, it is not evident to identify and implement all the rules for matching Arabic open domain questions.

On the other hand, we review related work about Arabic question classification methods that opted for learning-based approach:

- Al-Bayan (Abdelnasser et al., 2014): The system is designed to answer questions related to Quran. They used a new taxonomy based on Named Entity categories. The authors constructed 180 training questions and 50 testing questions according to their taxonomy. Using a SVM classifier, the overall accuracy with 3-folds cross validation was about 77.2%.
- Ahmed and Anto (2016) compared between SVM and Multinomial Naive Bayes classifiers. They used TF-IDF 1-gram and 2-gram features to train the classifiers. The dataset contains 300 questions for training and 200 questions for test. They obtained 100% on precision, 94% on recall and 97% on F1-measure.

3. Word representation

The learning-based approach involves a word transformation from its textual representation into a vector space model. This section reviews the word representation exploited in this work.

3.1. Term frequency-inverse document frequency (TF-IDF)

Several Arabic question classification methods (Abdelnasser et al., 2014; Ahmed and Anto, 2016) employed TF-IDF to represent questions into a vector space model. The TF-IDF weight of a word in a document d ($tfidf_t^d$) is computed following the equation:

$$tfidf_t^d = tf_t^d * \log\left(\frac{\#D}{\#D_t}\right) \tag{1}$$

where $tf_t^d, \#D$, and $\#D_t$ are respectively the frequency counts of word t in document d, the total number of documents, and the number of documents containing word t.

We can combine this representation with n-gram model which represents n words in one gram. For instance, if we take these two sentences "الكتابقرأ محمد" (Mohammed read the book) and "فتح محمد الكتاب" (Mohammed opened the book) the vocabulary of 1-grams (unigrams) is [[فتح - الكتاب - محمد - قرأ]]. The set of 2-grams (bigrams) is [[فتح محمد - محمد الكتاب - قرأ محمد]]. Each gram in the combined vocabulary of 1–2 grams, that is [الكتاب - محمد - قرأ – فتح محمد - محمد الكتاب - قرأ محمد - فتح], is used as a word in the Eq. 1 of TF-IDF. We called this representation the TF-IDF (1,2) grams. In similar manner, we can build TF-IDF (1,2,...,n) grams where n is a natural number and $n \geqslant 1$.

However, this representation suffers from a series of drawbacks: (1) the semantic of words, which is primordial to understand questions, can not be captured; (2) the size of the word representation is huge.

3.2. Enriched word vectors with subword information

In a major advance in 2013, Mikolov et al. (2013) and Mikolov et al. (2013) proposed a deep model for learning high-quality distributed vector representations of words. These representations capture a large number of both syntactic and semantic relationships between words. Several researches on Arabic text (El Mahdaouy et al., 2018; El Mahdaouy et al., 2016) applied word embedding model and proved its effectiveness. However, the shortcoming of the method proposed in Mikolov et al. (2013) is it ignores word morphology, since it assigns a distinct vector to each word. Therefore, languages with rich vocabularies and frequent rare words can not be represented properly. To overcome these limitations (Bojanowski et al., 2016) developed a new model that enriches word vectors with subword information. The authors represented each word as a sack of characters n-gram and they incorporated the word itself in the set of its n-gram. The word vector is the sum of its n-gram vectors. More formally, given a word w, its context word c, and d_w a dictionary that contains the set of n-grams appearing in w. Hence, the authors associate a vector representation r_d to each n-gram d, a word is represented by the sum of the vector representations of its n-grams. Therefore the vector representation of a word V_w is presented by the following equation:

$$V_w = \sum_{d \in d_w} \mathbf{r_d} \tag{2}$$

Let v_c the vector representation of the context word c. Then the model is trained to maximize the scoring function presented by the equation:

$$s(w, c) = \sum_{d \in d_w} \mathbf{r_d}^T \mathbf{v_c} \tag{3}$$

Fig. 2 illustrates the example of the word "مدينة" with all the 3-grams that represent the word.

In this work, we adopt the same model used in Bojanowski et al. (2016) as it is more suitable for Arabic language which has rich vocabulary and rare words. The model captures semantic relations between words and the dimension of the word vectors is 300.

4. Method

In this section, we present the proposed taxonomy extracted from Arabic language and our question classification method.

4.1. Taxonomy for arabic question classification

Several taxonomies related to QAS exist in the literature (Hao et al., 2015) and the choice of a taxonomy for question classification is not an obvious task. The language ambiguity is one of the main challenges that makes this choice so difficult. In order to reduce this ambiguity, the chosen taxonomy should maximize the distance in term of similarity between the question types.

Seeking to improve the Arabic question classification task, we built our Arabic taxonomy from deep and detailed Arabic interrogation rules.

Arabic language accepts 13 interrogation tools. Fig. 3 shows the Arabic interrogation tools. Our taxonomy, which is presented in Table 1, is extracted by analyzing the rules of Arabic interrogation tools and contains 7 types that take into account all the question types of the Arabic language. The following list presents for each Arabic taxonomy type its Buckwalter[5] transliteration, its description, and a question example:

- "**AlEAql**" – "العاقل" is the type of questions that ask which **person, people, organization**... The interrogative tool "mn" – "من" is generally used with this type of questions. For example

[5] http://www.qamus.org/transliteration.htm

Please cite this article as: A. Hamza, N. En-Nahnahi, K. A. Zidani et al., An arabic question classification method based on new taxonomy and continuous distributed representation of words, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.001

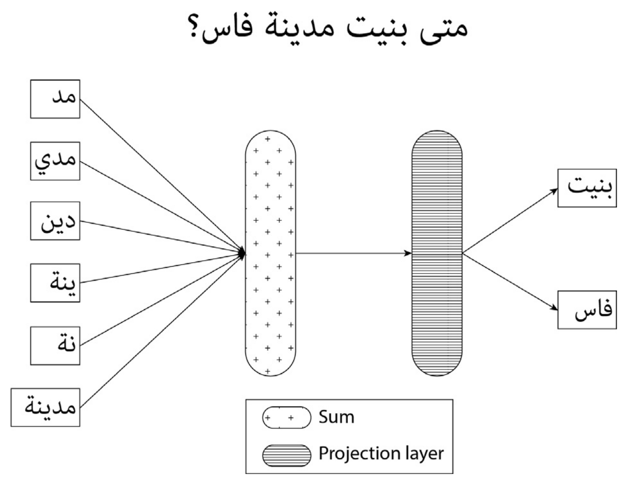

Fig. 2. Example of enriched word vectors with subword information.

"who is with you?" – من؟عندك the answer is "Mohamed" we cannot response with "a horse, a table, ...".

- "**gyr AlEAql > w Sfp EAql**" – "غير العاقل أو صفة عاقل" is the type of questions that ask about something that is not human like **animals, plants, inanimate, person characteristics, and others**. The tool "mA" – "ما" is used with this type of questions. For instance "what did you do?" "ما عملت؟".
- "**HAl Al$y' w hy}th**" – "حال الشيء و هيئته" is the type of questions that ask about **circumstance or structure of something**. The tool "kyf" – "كيف" is used with this types of questions. For instance "How are you?" "كيف أنت؟".
- "**AlmkAn**" – المكان is the type of questions that ask about **Location, country, citiy**... The tool ">yn" – "أين" is used with this type of questions. For instance "Where is Ali?" "أين علي؟".
- "**AlzmAn**" – الزمان is the type of questions that ask about **time**. The tool "mtY" – "متى" is used with this type of questions. For instance "When did you come?" "متى جئت؟".
- "**AlEdd**" – العدد is the type of questions that ask about **numbers**. The tool "km" – "كم" is used with this type of questions. For instance "How many pens have you?" "كم قلما عندك؟".
- "**AltSdyq**" – التصديق is the type of questions that ask for **yes or no** answers. The tool "hl" – "هل" is generally used with this type of questions. For instance "Did Mohammed come?" "هل حضر محمد؟".

The majority of taxonomies used for Arabic QAS, presented in Section 2, are based on Wh questions or built for restricted domain e.g., the taxonomy for the Holy Quran proposed by Abdelnasser et al. (2014). To our knowledge our taxonomy is the first taxonomy built from Arabic language linguistics studies (Cherif, 2007). For comparison purpose, we looked up for a taxonomy to compare our Arabic taxonomy. We chose the Li & Roth taxonomy, presented in Table 2, for the following reasons: (1) It is based on the semantic interpretation of question type; (2) It is the most widely used question taxonomy (Al Chalabi et al., 2015); (3) Its categories are shared by more taxonomies (Hao et al., 2015) which indicate the better universality. Despite the Li & Roth taxonomy can be used for questions written in Arabic language, this latter needs more appropriate and specific taxonomy.

4.2. Proposed question classification method

Our method is based on machine learning methods and continuous distributed representation of words. It takes into consideration morphological, syntactic and semantic relations between words. In addition, the dimension of obtained question vectors is reduced in comparison to those obtained by TF-IDF technique. The workflow of our method consists of three main steps: *(1)* Text

Table 1
Proposed Arabic Taxonomy.

Explanation	Class
Human, Group	العاقل
Entity, Animals,...	غير العاقل أو صفة عاقل
Status, Structure	حال الشيء و هيئته
Location	المكان
Time	الزمان
Numbers	العدد
Yes/No	التصديق

Table 2
Li & Roth Taxonomy (Li and Roth, 2002).

Coarse Classes	Fine Classes
ABBREVIATION	abbreviation, expression
DESCRIPTION	definition, description, manner, reason
ENTITY	animal, body, color, creative, currency, disease/medicine, event, food, instrument, lang, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUMAN	group, individual, title, description
LOCATION	city, country, mountain, other, state
NUMERIC	code, count, date, distance, money, order, other, period, percent, speed, temp, vol.size, weight

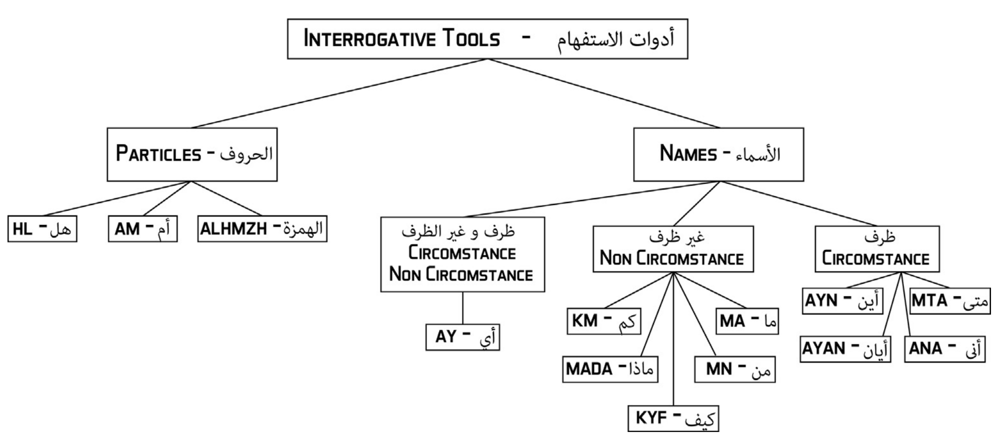

Fig. 3. Arabic interrogation tools.

Please cite this article as: A. Hamza, N. En-Nahnahi, K. A. Zidani et al., An arabic question classification method based on new taxonomy and continuous distributed representation of words, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.001

preprocessing *(2)* Question representation and *(3)* Question classification.

4.2.1. Preprocessing step

In this step, we start with a tokenization then we removed punctuations. We did not remove stop words since their ability to give meaning to a question and to affect significantly the classification result, e.g., the phrase "not consume pork" after stop words removal become "consume pork" which is the opposite of the original phrase.

4.2.2. Vector representation of question

We compute question vectors with two different methods: first, the most used word representation TF-IDF weighting with n-gram. Next, we adopt the representation explained in the Section 3.2 which represents a word as a bag of its character n-gram. A neural network is trained to maximize the score function that maps pairs of (word, context) into scores in $\mathbb{R}$ and computed as the dot product between the word vector and the context vector.

The obtained vectors are pre-trained on Wikipedia[6] and the dimension of each word vector is 300. Fig. 4 shows examples of semantic relations between words. A set of human professions are in the right while a set of animals are on the left, another key point to mention is that the feminine words are under their masculine words.

We represent a question as a bag of words. Therefore the question vector is calculated following the Eq. (4):

$$V_Q = \sum_i V_i \text{where} i \in D \tag{4}$$

where V_Q the question vector, D the set of words within the question Q, and V_i is a word vector where $i \in D$. For instance, the question "أين توجد مكة المكرمة؟" (Where is Makkah?) is represented by the sum of the word vectors of [[أين, توجد, مكة, المكرمة]]. A further point to mention is if a word did not exist in the training corpus then it could not be represented into the vector space model. Thus, the word is removed from the question. The obtained question vectors will be used as features in classification module.

4.2.3. Classification module

This module consists of classifying Arabic questions according to a specific taxonomy. We adopt a learning based approach that respects the following pipeline: First, we collect the raw questions with their expected types to accomplish the question classification task. Next, we apply a classification model to map a question, that is represented by a vector, to one and only one type of a well known taxonomy. At that point, we split the dataset into three parts including training set, validation set, and test set. Finally, we evaluate our classifier with different metrics.

5. Experimental results

To evaluate the effectiveness of our method, we collected 1302 Arabic questions from three data-sets TREC,[7] CLEF[8] and Moroccan school books. All collected questions are annotated according two taxonomies our Arabic taxonomy and Li & Roth taxonomy (Li and Roth, 2002) wich was enriched by Yes/No class to make a fair comparison between the two different taxonomies. We developed an application to alleviate the process of questions annotation.

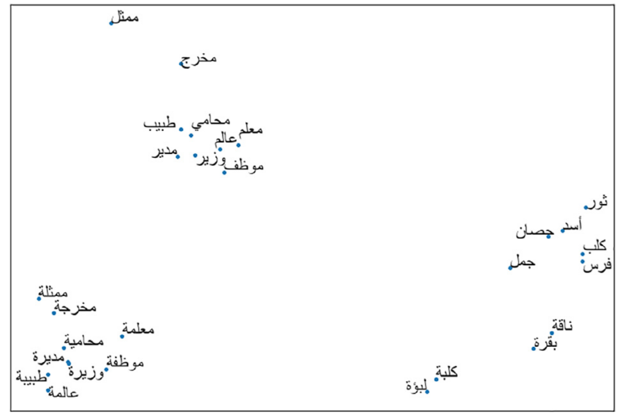

Fig. 4. Visualization 2D of semantic relations between words by projecting word vectors with PCA.

We performed two major evaluations to answer the following questions: (1) What is the question representation, between word embeddings and TF-IDF, that has the best performance on classifying Arabic questions? (2) Is the Arabic taxonomy better than Li & Roth taxonomy in the task of Arabic questions classification, and is there a relation between the classification model and the questions taxonomy?. In all experiments the dataset was split as follows 80% as train set and 20% as test set (train 1041–test 261). What's more, the measurements used to compare the performance of classifiers are: Accuracy, Precision, Recall, and F1-measure.

5.1. Performance evaluation of different question representations

In order to know the best question representation for Arabic question classification, we fixed the classifier and used different questions representations. We chose the SVM classifier which was used in several experiments (Abdelnasser et al., 2014; Ahmed and Anto, 2016). The classifier was trained on three datasets: (1) Arabic questions represented by the sum of word vectors which have a vector size equal to 300; (2) Arabic questions represented by 1-gram TF-IDF weighting, the obtained vector size is 2735; (3) Arabic questions represented by (1, 2)-gram TF-IDF weighting that creates a 5911 dimensional vector. Since the proportion of classes is not balanced (Table 3), we validate the model using 5-folds stratified cross validation. We performed grid search for hyper-parameters optimization. Table 4 shows that our method outperformed the existing TF-IDF based methods. The improvement is due to the use of word vectors representation (Bojanowski et al., 2016) that captures semantic relations and their size is very gainful.

5.2. Comparison between Arabic taxonomy and Li & Roth taxonomy

To evaluate the impact of our Arabic taxonomy, we compared the performance of several classifiers (SVM, MLP, XGBoost, Naive

Table 3
Arabic taxonomy distribution.

# of questions	% of questions	Class
122	9.37%	الزمان
241	18.51%	العاقل
123	9.44%	العدد
172	13.21%	المكان
115	8.83%	التصديق
56	4.30%	حال الشيء و هيئته
473	36.32%	غير العاقل أو صفة عاقل

[6] https://www.wikipedia.org/.
[7] http://trec.nist.gov/.
[8] http://www.clef-initiative.eu/.

Please cite this article as: A. Hamza, N. En-Nahnahi, K. A. Zidani et al., An arabic question classification method based on new taxonomy and continuous distributed representation of words, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.001

Table 4
Performance evaluation of different methods.

System	Metrics	Validation		Test	
		Our	Li & Roth	Our	Li & Roth
TF-IDF 1-gram + SVM	Accuracy	67.91%	68.76%	68.96%	70.11%
	Precision	–	–	75%	73%
	Recall	–	–	69%	70%
	F1-measure	–	–	68%	70%
TF-IDF (1,2)-grams + SVM	Accuracy	65.60%	66.36%	67.43%	65.9%
	Precision	–	–	76%	74%
	Recall	–	–	67%	66%
	F1-measure	–	–	67&	66%
Our method	Accuracy	86.5%	84.2%	90.03%	84.2%
	Precision	–	–	91%	85%
	Recall	–	–	90%	84%
	F1-measure	–	–	90%	84%

Table 5
Comparative results between Arabic taxonomy and Li&Roth taxonomy (Li and Roth, 2002) in order to classify Arabic questions, AR stand for Arabic taxonomy and Li for Li & Roth taxonomy.

Classifier & Taxonomy	Accuracy	Precision	Recall	F1
SVM_AR	90%	91%	90%	90%
SVM_LI	84.2%	85%	84%	84%
MLP_AR	87.3%	88%	87%	87%
MLP_LI	84.6%	85%	85%	85%
XGBoost_AR	83.1%	84%	83%	83%
XGBoost_LI	80.4%	82%	80%	80%
NB_AR	70.4%	73%	70%	71%
NB_LI	65.1%	69%	65%	65%
LR_AR	86.2%	87%	86%	86%
LR_LI	78.9%	80%	79%	78%

Bayes, Logistic Regression) applying the two taxonomies (Arabic taxonomy and Li & Roth taxonomy (Li and Roth, 2002). We used four classifiers in order to prove the independence between the proposed taxonomy and the model used for learning question classification. The obtained results, presented in Table 5, show that our Arabic taxonomy is more effective as compared to Li & Roth taxonomy(Li and Roth, 2002). The new taxonomy is inspired by Arabic language studies while the second taxonomy (Li and Roth, 2002) is made for english language. Also, we can notice that SVM classifier achieves the best performance. Indeed, Yang and Liu (1999) showed that Support Vector Machine(SVM) significantly outperform Neural Networks and Naive Bayes in text classification. Also, Ahmed and Anto (2016) found that SVM classifier outperforms Naive Bayes classifier in the task of Arabic question classification.

5.3. Error analysis

After the evaluation of the classifiers, we use confusion matrices (Fig. 5) to find the classes when the classifier fails the most.

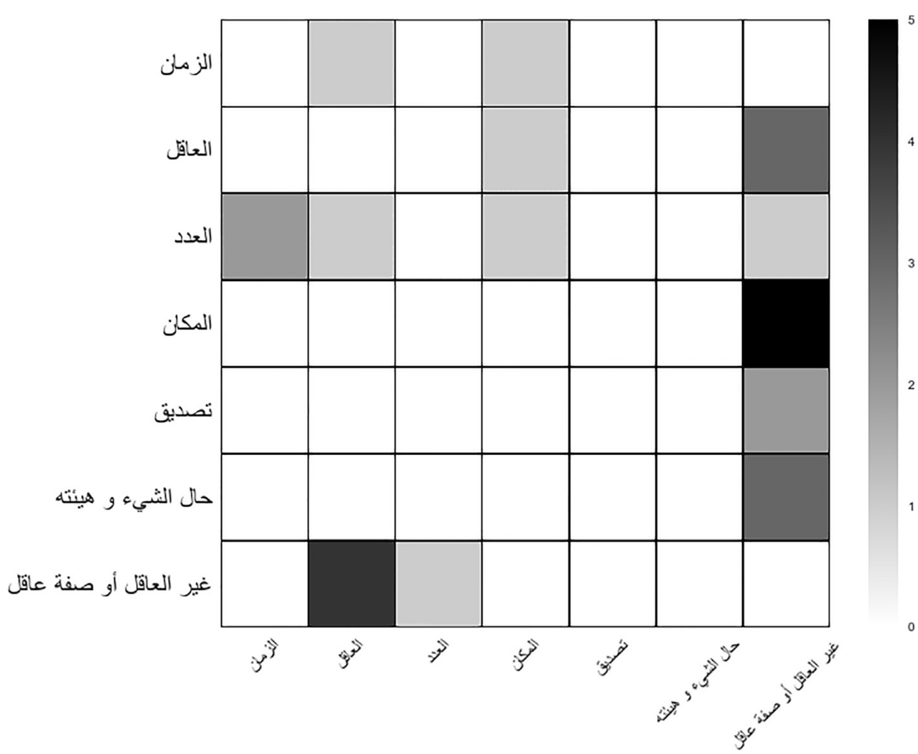

Fig. 5. Confusion matrices (diagonal filled with 0) of SVM classifier and Arabic taxonomy.

Please cite this article as: A. Hamza, N. En-Nahnahi, K. A. Zidani et al., An arabic question classification method based on new taxonomy and continuous distributed representation of words, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.001

Table 6
Examples of misclassified questions obtained by our method.

Predicted Arabic type	Expected Arabic type	Question
المكان	الزمان	في أية سنة تأسست منظمة حلف شمال الأطلسي ؟
AlmkAn	AlzmAn	In what year was North Atlantic Treaty Organization established?
المكان	العدد	ما هو ارتفاع سانت لويس بولاية ميسسوري ؟
AlmkAn	AlEdd	What is the height of St. Louis, Missouri?
غير العاقل	المكان	أين توجد الأنواع الكيميائية الطبيعية؟
gyr AlEAql	AlmkAn	Where are the natural chemical species found?
العدد	غير العاقل	ما هي المعادن التي لديها أعلى درجة انصهار ؟
AlEdd	gyr AlEAql	What metals have the highest melting point?

We analyze the errors made by the SVM classifier trained on questions labeled with Arabic taxonomy. Fig. 5 shows that the majority of errors is committed when predicting the class label "غير العاقل أو صفة عاقل" for questions that belong to "المكان" class. For instance, the question "أين توجد الأنواع الكيميائية؟" (Where are the natural chemical species found?) is hard to classify, even a human may have problems for classifying this question (Table 6).

The "ماذا يسمى داخل السفينة؟" (What is the ship inside called?) is an example where we can see that the expected label "Entity" has the second highest probability value 22% while the predicted label "Description" has 74% as probability value. Therefore, We can improve the performance of QAS by performing a multi label classification. To explain how this can be done, we take for example the two types of questions with the highest probabilities (in our case Description and Entity). These types are fed into the answer processing module, which will use this information to look up not for one type but for two types in the retrieved passages (by the passage retrieval module). Thus, the answer processing module can find the correct answer even if the question type with the highest probability was wrong.

6. Conclusion and future work

In this paper, we built a new Arabic taxonomy and proposed an Arabic question classification method based on continuous distributed representation of words and machine learning approach. We developed an application to annotate a set of questions, that are used to train our classifier, collected from TREC, CLEF and Moroccan school books. The proposed Arabic question classification method took advantages of both the quality of the question representation which is extracted using the word embeddings and the strength of machine learning methods. The proposed Arabic taxonomy, which is independent of the model used to classify Arabic questions, provides promising results in Arabic question classification compared to Li & Roth taxonomy. The obtained accuracy was 90% applying SVM classifier with our Arabic taxonomy. This work has revealed that our procedure is an enhancement of current Arabic question classification methods. The principal advantages are: The Arabic taxonomy is well suited for Arabic question classification task; The adopted word representation captures semantic and syntactic relations between words; The size of question representation is reduced compared to TF-IDF; Machine learning models are applied to classify question based on their vectors.

Future work will focus on building a full Arabic QAS by developing the modules that remain of the QAS: Document processing and Answer processing. We believe that word embeddings and machine learning methods will allow the creation of an efficient Arabic QAS.

References

Moldovan, D., Paşca, M., Harabagiu, S., Surdeanu, M., 2003. Performance issues and error analysis in an open-domain question answering system. ACM Trans. Inf. Syst. 21 (2), 133–154.
Mishra, A., Jain, S.K., 2016. A survey on question answering systems with classification. J. King Saud Univ.-Comput. Inf. Sci. 28 (3), 345–361.
Li, X., Roth, D., 2002. Learning question classifiers. In: Proceedings of the 19th International Conference on Computational Linguistics – Volume 1. Association for Computational Linguistics, pp. 1–7.
Sarrouti, M., El Alaoui, S.O., 2017. A machine learning-based method for question type classification in biomedical question answering. Methods Inf. Med. 56 (03), 209–216.
Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.
Mohammed, F., Nasser, K., Harb, H., 1993. A knowledge based arabic question answering system (aqas). ACM SIGART Bull. 4 (4), 21–30.
Hammo, B., Abu-Salem, H., Lytinen, S., 2002. Qarab: a question answering system to support the arabic language. In: Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages. Association for Computational Linguistics, pp. 1–11.
Salton G. The Smart Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs.
Benajiba, Y., Rosso, P., Lyhyaoui, A., 2007. Implementation of the arabiqa question answering system's components. In: Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, pp. 3–5.
Trigui, O., Belguith, L.H., Rosso, P., 2010. Defarabicqa: arabic definition question answering system. In: Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, pp. 40–45. Valletta, Malta.
Abouenour, L., Bouzoubaa, K., Rosso, P., 2012. Idraaq: new arabic question answering system based on query expansion and passage retrieval. In: CLEF (Online Working Notes/Labs/Workshop).
Al Chalabi, H.M., Ray, S.K., Shaalan, K., 2015. Question classification for arabic question answering systems. In: Information and Communication Technology Research (ICTRC), 2015 International Conference on. IEEE, pp. 310–313.
Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N., Torki, M., 2014. Al-bayan: an arabic question answering system for the holy quran. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 57–64.
Ahmed, W., Anto, B., 2016. Classification of arabic questions using multinomial naive bayes. Int. J. Latest Trends Eng. Technol., 82–86 (Special Issue SACAIM)
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst., 3111–3119
Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
El Mahdaouy, A., El Alaoui, S.O., Gaussier, E., 2018. Improving arabic information retrieval using word embedding similarities. Int. J. Speech Technol. 21 (1), 121–136. https://doi.org/10.1007/s10772-018-9492-y.
El Mahdaouy, A., Gaussier, E., El Alaoui, S.O., 2016. Arabic text classification based on word and document embeddings. In: International Conference on Advanced Intelligent Systems and Informatics. Springer, pp. 32–41.
Hao, T., Xie, W., Chen, C., Shen, Y., 2015. Systematic comparison of question target classification taxonomies towards question answering. In: Chinese National Conference on Social Media Processing. Springer, pp. 131–143.
Cherif, M.I.M., 2007. Methods of questioning in the rhetorical search and its secrets in the holy quran (Ph.D. thesis). International Islamic University, Islamabad.
Yang, Y., Liu, X., 1999. A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval ACM, pp. 42–49.

Please cite this article as: A. Hamza, N. En-Nahnahi, K. A. Zidani et al., An arabic question classification method based on new taxonomy and continuous distributed representation of words, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.001