International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 82-86
e-ISSN:2278-621X

CLASSIFICATION OF ARABIC QUESTIONS USING MULTINOMIAL NAIVE BAYES AND SUPPORT VECTOR MACHINES

Waheeb Ahmed[1] and Babu Anto P[2]

Abstract- Question classification plays a very important role in Question Answering systems. It gives a label to a question depending on the type of the question. This label will be used by the Answer Extraction module to extract the correct answer. Since there are variety of Natural Language Questions, the task of classifying different questions becomes hard and challenging. Very limited research has been done on classifying Arabic Questions using Machine Learning Techniques. In this paper, we used Support Vector Machines(SVM) and Multinomial Naive Bayes(MNB) to classify Questions. The types of questions classified includes Who, What, Where, When, How many, How much, How and Why. The labels that will be given to these questions respectively are Person/Definition, Location, Time/date, Number/Count, Quantity, Manner and Reason. The SVM showed higher accurate results than MNB. The dataset consisted of 300 questions from the Arabic Wikipedia. The precision of both the SVM and the MNB is equivalent to precision of 1. The achieved F1 measure for SVM is .97 and for the MNB is .95 which is a promising result.

Keywords – Question Classification, Question Answering, Machine Learning.

I. INTRODUCTION

Question Answering is a computer science discipline that uses Information Retrieval(IR) and techniques of Natural Language Processing (NLP) to answer questions posed by humans to get the proper answer[1]. Question Answering Systems have two domains: Open domain and Closed domain. Open domain QA deals with everything whereas closed domain deals with questions related to a specific domain(Quran, Medical Applications, Biology etc)[2]. Our work focuses on Closed domain ,that is, the Arabic Wikipedia. Classifying questions posed by users in a question answering is considered a very challenging problem [3]. Question classification in QA is a crucial step, because it will help in anticipating the type of answer and this will narrow down the search space for finding the correct answer. The purpose is to concentrate the answer extraction task only on those text segments related to the expected type of answer which is identified by the question classification module[4].

There are two approaches for classifying questions: one is rule-based approach and the other is machine learning approach. Recently, supervised machine learning techniques are adopted, which train a classifier from examples that are manually annotated (questions along with their corresponding answer types). In fact, creating a training and testing set is a time-consuming process, but no rule-writing skills are required[5]. Hence, we used the machine learning approach by training a classifier on a set of questions derived from Arabic Wikipedia.

[1] *Department of Information Technology Kannur University, Kannur, Kerala, India*
[2] *Department of Information Technology Kannur University, Kannur, Kerala, India*

II.RELATED WORK

Al Chalabi[6] proposed question classification methods for Arabic questions using regular expressions and context free grammars. They used Nooj Platform[7] to write regular expressions and used linguistic patterns to identify the type of expected answer.

Ali[8] proposed a question classification using support vector machines. Their classifier can classify only three types of questions namely "Who", "Where" and "What". They used 1-gram, 2-gram ,3-gram features and TF-Weighting and they indicated that the 2-gram feature produced the best classification with a performance of 87.25% using F1 measure.

Abdenasser[9] used SVM classifier to classify Quranic questions and they got an overall accuracy of the classifier equivalent to 77.2%. Their data set consisted of 230 questions from Quranic domain.

III.PROPOSED METHODOLOGY

SVM is a machine learning technique that classifies text. It proved to be an efficient classifier for text categorization. MNB is an advance version of Naive Bayes that is designed for classifying text documents. It gets the words counts in documents rather than the presence and absence of particular words as traditional Naive Bayes does[10]. We are using support vector machines and Multinomial Naive Bayes to classify the given question according to the training data that we built. The training data set consists of 300 questions derived from the Arabic Wikipedia. The testing data set consists of 200 questions which are translated from Text Retrieval Conference(TREC 10)[1]. We used 1-gram and 2-gram features while training the classifier.

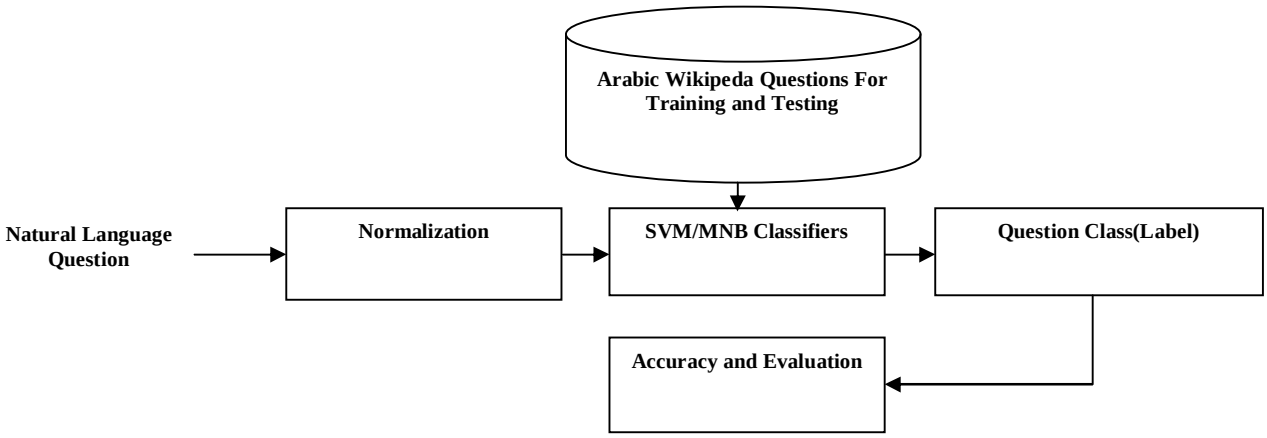

Figure 1. Question Classification Using SVM and MNB

A natural language question is given by the user. The Arabic diacritics(Vowels) will be removed and the normalized text will be given to the classifiers. The classified question will be given a label(The labels are provided in Table 1 in the next section). The accuracy and evaluation stage is used to evaluate the performance of the classifiers.

IV.QUESTION TYPE TAXONOMIES

Questions are classified into different types: (Who)من , (When)متى , (Where)أين , (What)ماهو-ما هي , (How many)كم عدد , (How much)كم كمية , (How)كيف , (Why)لماذا .

Table -1 Classes of Questions

Question type	Expected Answer type (LABEL)	Examples
(who)من	(Person)شخص	man-hua من هو/man-hya من هي : Questions that starts with Who(نم) asks for a person name, so the class label given to this question is Person(صخش). So the answer expected for this type is a person name. e.g: Who is the president of the United States? من هو رئيس الولايات المتحدة؟
(where)أين	(Location)مكان	ayin-اين: This question has the meaning of 'Where'. It looks for answer of the type Location, Location is further divided into four subclasses which includes City(), State, Country, and Other. e.g: Where is London?. The main class for this question is Location. The subclass is City.
(when)متى	(Time)زمان	mata-ىتم: This kind of questions asks for Time/Date. So the main class is Number(مقر) and the subclass is Time(تقو) or Date(خيرات). e.g: When did Tunisia gained independence? ىتم ؟ سنوت تلقتسا
(how much)كم	(Quantity)كمية	Kam-kamyat كم كمية: This question asks for Quantity. e.g: how much blood in human body? كم كمية الدم في جسم الانسان؟
Question Type	Expected Answer Type (LABEL)	Examples
(how many)كم	(Count)عدد	Kam-Aded كم عدد : This is equivalent to ' How Many'. The main class for this question is Number and the subclass is Count. How many continents are there? كم (عدد القارات؟)
(what)ماهو-ما هي	(Thing)شيء	This question asks for entity. e.g: What is the the color of the sun? (ما هو لون الشمس ؟) In this case , the class of this question will be Entity.
	(Definition) فيرعت	ma-hua ما هو/ma-hya ما هي : It asks for definition. Like What is/are in English. The class of this question is Definition(تعريف). e.g: What is Computer?(ما هو الكمبيوتر ؟)
(How)كيف	(Manner)الوسيلة	kaif- كيف : This question is asking for the manner(How). It is given label 'Manner'. e.g: How water can be transferred from liquid to solid? (كيف يمكن تحويل الماء من الحالة السائلة الى الحالة الغازية؟)
(Why)لماذا	(Reason)المبرر	limatha- لماذا : This question is asking for reason. So it is given the label Reason. Why do birds sing? (لماذا تغني الطيور؟)

V.PERFORMANCE EVALUATION OF QUESTION CLASSIFIERS

To measure the performance of a question classifier we use precision and recall of the system. Precision (P) is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP).

$$P = \frac{T_P}{T_P + F_P}$$

Recall (R) is defined as the number of true positives (TP) over the number of true positives plus the number of false negatives (Fn).

$$R = \frac{T_P}{T_P + F_n}$$

Where True Positive (T_P) is the set of questions that is correctly assigned to the class , False Positive (F_P) is the set of questions that are incorrectly assigned to the class, False Negative (F_n) is the set of questions that are incorrectly not assigned to the class, and True Negative (Tn) is the set of questions that are correctly not assigned to the class.

These Precision and Recall are also related to the (F1) score, which is defined as the harmonic mean of precision and recall.

$$F1 = 2\frac{P \times R}{P + R}$$

A number of 200 that are used to test MNB and SVM are classified correctly with a precision of 1.

Table -1 Performance Evaluation For MNB

Question Type	Precision	Recall	F1-measure
(who)من	1	.96	.97
(where)أين	1	.94	.96
(when)متى	1	.90	.94
(how much)كم كمية	1	.89	.94
(how many)كم عدد	1	.93	.96
(what)ماهو-ما هي	1	.87	.93
(How)كيف	1	.91	.95
(Why)لماذا	1	.88	.93
AVG	1	.91	.95

Table 1 show the Precision, Recall and F-measure for the listed question types obtained MNB for classifying the questions. The obtained average precision by MNB is 1, the recall is .91 and the F1-measure is .95.

Table -2 Performance Evaluation For SVM

Question Type	Precision	Recall	F1-measure
(who)من	1	.98	.99
(where)أين	1	.95	.97
(when)متى	1	.96	.98
(how much)كم كمية	1	.96	.98

(how many) كم عدد	1	.93	.96
(what)ماهو-ما هي	1	.92	.95
(How)كيف	1	.90	.94
(Why)لماذا	1	.92	.95
AVG	1	.94	.97

Table 2 show the Precision, Recall and F-measure for the listed question types obtained by SVM for classifying the questions. The obtained average precision by SVM is 1, the Recall is .94 and the F1-measure is .97. The result is greatly promising comparing to some recent research on Question Answering of English language . Systems with recall 0.63 and precision 0.7 [11] and recall 0.73 and precision 0.73 [12] . Hence, the results that we got shows the effectiveness of Support Vector Machines and Multinomial Naive Bayes in classifying the questions.

VI.CONCLUSION

In this paper, we proposed question classification method using SVM and MNB for Arabic questions. We trained the classifiers on 300 questions derived from the Arabic Wikipedia and tested them using a set of 200 translated questions from TREC. The results are very promising and can be used in developing Arabic Question Answering Systems.

REFERENCES

[1] E. M. Voorhees," Overview of the TREC 2001 question answering track," Proceedings of the 10th Text Retrieval Conference, pp. 42–52, 2001.

[2] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," International Journal of Research and Reviews in Information Sciences (IJRRIS) , vol. 2, no. 3, pp. 211-221, 2012.

[3] V. Punyakanok, D. Roth, and W. tau Yih, "Natural language inference via dependency tree mapping: An application to question answering", Computational Linguistics, vol. 6, no. 9, 2004.

[4] Antonio, Claudia, Manuel and Luis, "Using Machine Learning and Text Mining in Question Answering, Evaluation of Multilingual and Multi-modal Information Retrieval", Volume 4730 of the series Lecture Notes in Computer Science, pp 415-423, 2007.

[5] Oleksander and Marie-Francine, "A survey on question answering technology from an information retrieval perspective", Information Sciences 181, pp. 5412-5434, 2011.

[6] H. M. Al Chalabi, "Question Classification for Arabic Question Answering System," International Conference on Information and Communication Technology Research (ICTRC), pp. 310 – 313, 2015.

[7] Nooj website:http://www.nooj4nlp.net-Last visited-September, 2016.

[8] Ali Muttalib, Lailatul Qadri. Question Classification using Support Vector Machine And Pattern Matching. Journal of Theoretical and Applied Information Technology, E-ISSN:1817-3195, 2016.

[9] Heba Abdelnasser, Reham Mohammed, Al-Bayan: An Arabic Question Answering System for the Holy Quran. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pp. 57–64, 2014.

[10] McCallum and k. Nigam, A Comparison of Event Models for Naive Bayes Text Classi cation, In proceedings of the AAAI/ICML-98 on Learning For Text Categorization, AAAI Press, pp. 41-48,1998.

[11] Borhan Samei, Haiying Li, Fazel Keshtkar, Vasile Rus, and Arthur C. Graesser. Context-based speech act classification in intelligent tutoring systems. In Intelligent Tutoring Systems,Springer International Publishing, pp. 236-241, 2014.

[12] Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga, Elena Cabrio, Philipp Cirniano, Sebastian Walter. Question Ansering over Linked Data (QALD-4). In Working Notes for CLEF 2014 Conference, volume 1180 of CEUR Workshop Proceedings, pp. 1172–1180, 2014.