International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Malayalam Questions Classification in Question Answering Systems using Support Vector Machine

Bibin P.A[1]*, Babu Anto P[2]

[1] Department of Computer Science, St. Pius X College, Rajapuram, Kerala, India
[2] Department of Information Technology, Kannur University, Kerala, India

**Corresponding author: bibinpa@gmail.com, Tel: +91 9447877128*

Available online at: www.ijcseonline.org

Accepted: 26/Jan/2019, Published: 31/Jan/2019

***Abstract*—** We consider Question answering systems (QAS) as the next step in information retrieval, allowing users to create questions in natural language and get concise answers. Researches show that exact classification of questions with respect to the expected answer type is imperative to make a successful QAS. The duty of classifying distinctive questions becomes hard and challenging because there are variety of Natural Language Questions. Due to the agglutinative nature researchers find so many difficulties in Malayalam based QAS. So a very limited researches have been done in classifying Malayalam Questions with the help of Machine Learning Techniques. In this paper, we have used Support Vector Machines (SVM) to classify Questions. In Malayalam we can classify the question into following types എന്ത് (what), എപ്പോൾ (when), എങ്ങനെ (how), എവിടെ(where), എന്തുകൊണ്ട്(why), എത്ര (how many/how much) and ആര്(who). For Malayalam Question classification using SVM 1is the average precision, 0.93 is the average recall and the average F1 Score is 0.95. So the outcome that we obtained shows the effectiveness of Support Vector Machines in classifying the question.

***Keywords*—** Malayalam Question Classification, Support Vector Machine, Machine Learning, Question Answering

I. INTRODUCTION

With the invention of computers the life of human has become much easier. In this century with the help of any web search engines any person can access any data, which is at any place of the world at his/her a finger tip. We expect the computers to act like an intelligent human being. Because of this a new stream called Artificial Intelligence (AI) has emerged in the field of Computer Science.

Natural Language Processing (NLP) is a branch of AI, where a user is interacting in his own language such as English, Malayalam etc to an intelligent system. Since there are many languages in the world, understanding of Natural Language become an important hurdle in making the computers intelligent. Question Answering is one of the major tasks in NLP. Question Answering in natural languages can be done using various techniques in NLP such as Information Retrieval (IR), Machine Learning, Knowledge Representation, etc [1],[2],[3].

Based on the application domain the Question Answering System (QAS) can be classified into: Restricted domain QAS and General domain QAS. Restricted domain QAS answers only domain specific questions [4]. Here the answers will be searched only within a specific document collection. General domain QAS answers questions from all domains, but the answer will not be very precise [5]. In our research we are mainly concentrating on the Restricted domain, which is tourism.

One of the most challenging problem in question answering is to classify the question which is given by the user [6]. A Question Classification module has two main advantages. 1) It provides an outlook on the type of the answer which helps us to proceed further to find out and confirm the answer. 2) It gives us the information that help downstream processes in finding answer selection strategies that may be answer type specific, rather than uniform.

Usually, Question Classification can be achieved by two approaches: Rule-based approach and machine learning approach. In this paper with the help of machine learning approaches we are deriving the expected answer types. This work contains three parts: (i) A scheme of classification of answer types into which questions should be arranged (ii) With the correct answer type scheme of classification a corpus of questions are to be disposed and (iii) an algorithm that gain knowledge from this corpus and makes the correct prediction [7]. With the help of supervised machine learning techniques, we could train a classifier using the corpuses that are manually annotated with their question and its corresponding answer type. Creating corpus for training and testing is a much time-consuming and a tedious task, but the advantage is that we do not require any rule-writing skills [8].

Section 1 contains the introduction of NLP and QAS, Section 2 contains the motivation for QAS and importance of Malayalam language, Section 3 contain the related work of QAS, Section 4 contain the architecture and essential steps of Question Classification, section 5 describes Effectiveness and Efficiency, Section 6 describes results and discussion of question classification and Section 7 concludes research work.

II. MOTIVATION FOR QUESTION ANSWERING SYSTEM(QAS)

From the very beginning itself, human beings are always in search of information. With the aid of web search engines or other IR techniques appropriate information is available to everyone at their finger tips. Question answering can be considered as a specialized type of IR. When we are using web search engine, we are getting only the relevant pages but we are interested in getting the precise answer to the questions. Question answering helps us to obtain the exact answer, and it includes NLP, IR, Machine Learning (ML), Knowledge Representation, Logic and Inference, Semantic Search, etc. So we can say that almost every branches of AI is contained in Question Answering. QAS can be used in any domain, such as tourism, teaching, personal assistants, medical science, etc. It can be used in every situation where we are in need of help from the computers. So this research is very much relevant in the current scenario.

In India there are 22 scheduled languages and many more unscheduled languages. Malayalam is one among the scheduled languages. Malayalam belongs to the Dravidian family of languages and is the official language of south Indian state Kerala. Malayalam Language is rich in Morphological inflections ie, adding of suffixes, prefixes and infixes to the root or the stem word. Due to its agglutinative nature, researchers find so many difficulties in Malayalam based QAS. The literacy rate of Kerala is 96.7 percentages, which is the highest literacy rate in India. Most of the Keralites knows only Malayalam language and are not so good in using English language. Kerala is known as God’s own Country. Kerala is famous for its Ayurvedic treatments, high mountains, gorges and deep-cut valleys, lush and evergreen rain forests, coconut palms, backwaters, and food items. Keralites also want to know a lot of things about the different tourism spots and its related things. Getting exact answer from a set of documents in Malayalam for a particular question is very difficult. They required a system that can help them to find a precise or short answer to their questions.

III. REVIEW OF LITERATURE

Nowadays to get the correct answer to a question from internet with the help of a web browser using own language we are in need of QAS. L. Hirschman *et al.* gives us a brief idea of the upcoming research trends in QA in English. The background of the TREC QA evaluations, the results obtained from it, the methodology used for evaluation, the four important methods which are most important in QA, etc. are discussed in this paper [9].

Dell Zhang *et al.* had done a comparison between the different algorithms used for Machine Learning. The algorithms which they used were Support Vector Machine (SVM), Nearest Neighbours (NN), Naïve Bayes (NB), Decision Tree (DT) and Sparse Network of Winnows (SNoW). They found that SVM can achieve performance improvement over other Machine Learning algorithms [10].

Bhoir V *et al.* proposed solution of QAS works for a specific domain of tourism. The crawler developed in the system gathers web page information which is processed using NLP and procedure programming for a specific keyword. The system returns precise short string answers or list to natural language questions related to tourism domain like distance, person, date, list of hotels, list of forts, etc [11].

Pragisha K. *et al.* given the system which finds answers of Malayalam factual questions by analyzing a repository of Malayalam documents for handling the four classes of factual questions in Malayalam for closed domain. The QA system is divided into three modules as Question Analysis, Text Retrieval and answer snippet extraction and Answer identification [12].

Raji Sukumar *et al.* presented a model for developing intelligent query processing in Malayalam. For this they had selected a time enquiry system in Malayalam language. Natural Language Query Processing System is a restricted domain system, deals with the natural Language Queries on time enquiry for different modes of transportation. The system performs a shallow syntactic and semantic analysis of the input query. After the knowledge level understanding of the query, the system triggers a reasoning process to determine the type of query and the result slots that are required [13].

IV. PROPOSED MALAYALAM QUESTION CLASSIFICATION

SVM is a supervised machine learning method that examines the given data and arrange them into one of the two categories. It is a clever method to overcome over fitting and can deal with a large number of features with less computation. Using SVM we are trying to find out different question keywords in Malayalam from our question phrase and then classify them based on the training data which we have already created. We have used testing data sets which are translated from Text REtrievial Conference (TREC 10) and then assigned labels to them.

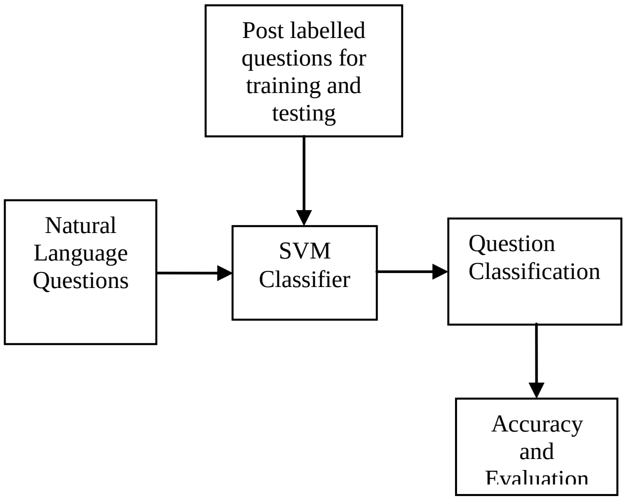
Post labelled questions for training and testing
Natural Language Questions
SVM Classifier
Question Classification
Accuracy and Evaluation

Figure 1: Question Classification using SVM

In Malayalam we can classify the question into the following types: എന്ത് (what), എപ്പോൾ (when), എങ്ങനെ (how), എവിടെ (where), എന്തുകൊണ്ട് (why), എത്ര (how many/how much) and ആര് (who).The following table shows the different types of questions and their assigned labels with examples.

Table 1: Question type

Question Type	*Expected Answer Type*	*Example and label*
എന്ത് (what)	കാര്യം (Thing)	Eg. ബോട്ടിൽ എന്ത് ഇന്ധനമാണ് ഉപയോഗിക്കുന്നത്? (What fuel does the boat use?) *For this question the main class is* ENTITY *and sub class is* Substance
	നിർവ്വചനം *(Definition)*	Eg. എന്താണ് ടൂറിസം? (what is tourism) *For this question the main class is* DESCRIPTION *and sub class is* Definition
എപ്പോൾ(when)	സമയം *(time)*	*Eg.* എപ്പോൾപ്രദർശനംആരംഭിക്കും? (*When does the show start?)* *For this question the main class is* NUMERIC *and sub class is* Date/Time
എങ്ങനെ(how)	വിധത്തിൽ *(manner)*	Eg. എനിക്ക് മൂന്നാറിൽ എങ്ങിനെ എത്തിച്ചേരാം? (How can I reach Munnar?) *For this question the main class is* DESCRIPTION *and sub class is Manner*
എവിടെ(where)	സ്ഥലം *(Location)*	Eg. തജ്മഹൽ എവിടെയാണ്? (Where is taj mahal?) *For this question the main class is* LOCATION *and sub class is City or* Country
എന്തുകൊണ്ട് (why)	കാരണം *(Reason)*	Eg. നിങ്ങൾഎന്തിനാണ്കണ്ണൂരിലേക്ക് പോകുന്നത്? (Why are you going to Kannur?) *For this question the main class is* DESCRIPTION *and sub class is* Reason

എത്ര (how many / how much)	എണ്ണം *(Count)*/അളവ് (Quantity)	Eg. ഇൻഡ്യയിൽ എത്ര സംസ്ഥാനങ്ങൾ ഉണ്ട്? (How many states are there in India?) *For this question the main class is* NUMERIC *and sub class is* Count
ആര്(who)	വ്യക്തി *(Person)*	Eg. കേരളത്തിന്റെ ടൂറിസം മന്ത്രി ആര്? (Who is the tourism minister of Kerala?) *For this question the main class is* HUMAN *and sub class is* Individual

V. EFFECTIVENESS AND EFFICIENCY

In our Malayalam Question classifier system, the user will provide a natural language question to the system. The Malayalam Question classifier will train the system based on the 200 training data set which we had already labelled. Then it will test it using the testing data set. Based on this, the question posed will be classified and will be assigned a label. The effectiveness and efficiency of the system is measured with the help of precision and recall system.

Precision can be defined as in what ratio we could predict positive identifications was actually correct? Precision (P) is the total count of True Positives (TP) divided by the total count of True Positives (TP) plus the total count of False Positives (FP).

$$P = \frac{TP}{TP + FP} \qquad (1)$$

Recall can be defined as in what ratio we could predict actual positives was identified correctly? Recall (R) is the total count of True Positives (TP) divided by the total count of True Positives (TP) plus the total count of False Negatives (FN).

$$R = \frac{TP}{TP + FN} \qquad (2)$$

A True Positive (TP) is a result where the system correctly forecast the positive class. A True Negative (TN) is a result where the system correctly forecast the negative class. A False Positive (FP) is a result where the system incorrectly forecast the positive class. A False Negative (FN) is a result where the system incorrectly forecast the negative class. In Binary Classification, we can have either positive class or negative class. Positive class denotes the object we are searching and the negative class denotes the other chance.

F1 score or F Measure is required when we are in need to find a balance between Precision and Recall. F1 score can be defined as the weighted average of Precision and Recall.

$$\text{F1 Score} = 2 \, \frac{P * R}{P + R} \qquad (3)$$

Various 200 questions for training and 100 questions for testing are utilized with SVM and classified effectively with precision of 1.

Table 2: Performance Evaluation for SVM

Question Type	Precision	Recall	F1 Score
എന്ത് (what)	1	0.92	0.95
എപ്പോൾ (when)	1	0.95	0.97
എങ്ങനെ (how)	1	0.89	0.94
എവിടെ (where)	1	0.94	0.96
എന്തുകൊണ്ട് (why)	1	0.91	0.95
എത്ര (how many/how much)	1	0.94	0.96
ആര് (who)	1	0.96	0.97
Average	1	0.93	0.95

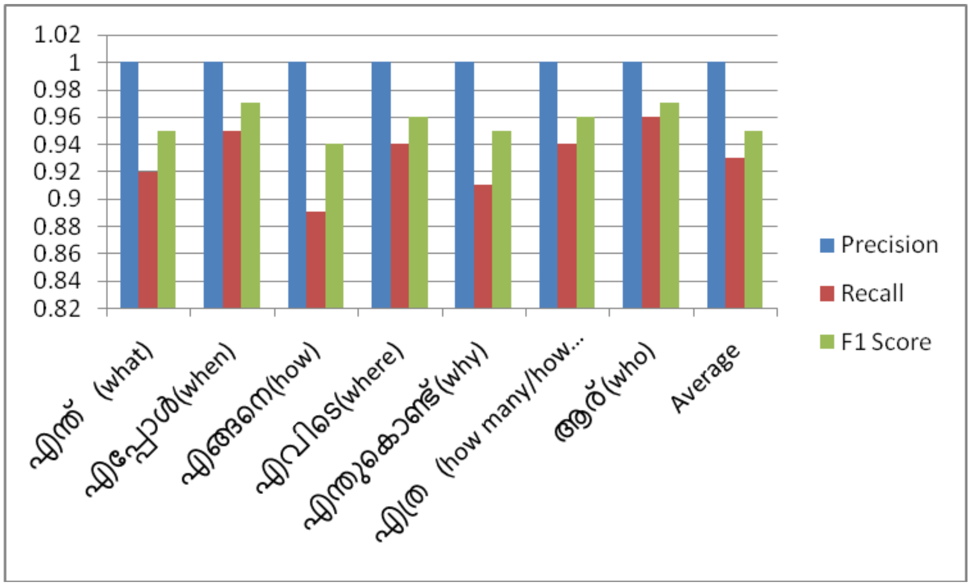

Figure 2: Graphical Performance Evaluation for SVM

VI. Results and Discussion

The precision, recall and F1 score for the recorded question types obtained by SVM for classifying the questions are demonstrated on table 2 and figure 2. The obtained average precision by SVM is 1 the recall is 0.93 and the F1 Score is 0.95. The outcome is extraordinarily encouraging compared to some current research on QAS of English language with average precision 0.7 and recall 0.63 [14] and precision 0.73 and recall 0.73 [15]. Consequently, the result we obtained shows the effectiveness of SVM in classifying the questions.

VII. CONCLUSION

For Question classification, average precision, recall and F1 score are 1, 0.93 and 0.95 respectively. Hence, the result that we obtained shows the effectives of Support Vector Machines in classifying the question. The Question classification has an important role in determining the techniques used for extracting the correct answer. So we can expect that with the help of Question Classification method we could develop a Malayalam QAS with much better accuracy and precision.

REFERENCE

[1] V. Lopez, V. Uren, M. Sabou and E. Motta, "*Is question answering fit for the Semantic Web? A survey* " , Semantic Web, Vol. **2**, Issue. **2**, pp. **125-155**, **2011**.

[2] S. K. Dwivedia, V. Singhb, "*Research and reviews in question answering system*", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), **India**, pp.**417-424**, **2013**

[3] G. Suresh kumar, G. Zayaraz, "*Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems*", Journal of King Saud University - Computer and Information Sciences, Vol. **27**, Issue. **1**, pp. **13-24**, **2015**.

[4] D. Mollá, J.L. Vicedo "*Question answering in restricted domains: an overview* ", Computational Linguistics, Vol. **33**, issue. **1**, pp. **41-61**, **2007**.

[5] N. Indurkhya, F.J. Damereau., "*Handbook of Natural Language Processing (second ed.)*" , Chapman & Hall/CRC, Boca **Raton** , **2010**.

[6] V. Punyakanok, D. Roth, and W.-t. Yih., "*Natural language inference via dependency tree mapping: An application to question answering*", Computational Linguistics, Vol. **6**, Issue. **9** pp. **1-10**, **2004** .

[7] H Sundblad, "*Question Classification in Question Answering systems*", Phd Thesis Submitted to Department of Computer and information Science at **Linkoping University**, **2007**.

[8] O. Kolomiyets , M.F. Moens., "*A survey on question answering technology from an information retrieval perspective*", Information Sciences—Informatics and Computer Science, Intelligent Systems, Applications: An International Journal, Vol.**181**, Issue.**24**, pp. **5412-5434**, **2011**

[9] L. Hirschman, R. Gaizauskas, " *Natural language question answering: The view from here*", Natural Language Engineering, Vol.**7**, Issue.**4**, pp. **275-300**, **2001**.

[10] D.Zhang, W.S. Lee, "*Question Classification Using Support Vector Machines*", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, Toronto, **Canada**, pp.**26-32**, **2003**.

[11] V. Bhoir, M. A. Potey, "Question answering system: A heuristic approach," *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, **Bangalore**, pp. **165-170**, **2014**.

[12] K. Pragisha, P. C. Reghuraj, "*A Natural Language Question Answering System in Malayalam Using Domain Dependent Document Collection as Repository",* International Journal of Computational Linguistics and Natural Language Processing, Vol. **3** Issue. **3**, pp.**534-539**, **2014**.

[13] S. A. Raji , B. P. Anto , " *Intelligent Query Processing In Malayalam*", International Journal on Computational Sciences & Applications (IJCSA), Vol.**3**, Issue.**2** pp. **51-59**, **2013**.

[14] B. Samei, H. Li, F. Keshtkar, V. Rus, A. C. Graesse, "*Context-Based Speech Act Classification in Intelligent Tutoring Systems*", Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science, Vol.**8474**, pp. **236-241**, **2014**

[15] C. Unger, C. Forascu, V. Lopez, A.C. Ngonga Ngomo, E. Cabrio, P. Cimiano, S. Walter, " *Question Answering over Linked Data (QALD-4)*" , In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), Working Notes for CLEF 2014 Conference Sheffield, **United Kingdom** pp. **1172-1180**, **2014**.

Authors Profile

Bibin P A pursued Bachelor and Master of Computer Application from Marian College Kuttikkanam, Mahatma Gandhi University, Kottayam, Kerala, India in 2004 and 2007. He is currently pursuing Ph.D. from Kannur University, Kerala, India and currently working as Assistant Professor in Department of Computer Science, St. Pius X College, Rajapuram, Kannur University, Kasargod, Kerala since 2011. His main research work focuses on Machine Learning, Natural Language Processing, Data Mining, Information Reterival and Artificial Intelligence. He has 8 years of teaching experience and 3 years of Research Experience.

Babu Anto P pursued Bachelor of Science from Calicut University in 1980. 2. Completed Master of Science from Cochin University in 1982. 3. Awarded Ph.D. from Cochin University of Science And Technology in 1992. Currently working as an Assosiate Professor in Department of Information Technology, Kannur University ,Kerala, India from 2003 onwards. He is a life member of CSI, ISTE, ASI, International Association of Computer Science and Information Technology and International Association of Engineers. He has published more than 75 research papers in reputed international and national jouranals. Under his supervision and guidance 6 Ph.D. degree were awarded and 5 are pursuing Ph.D. His main research work focuses on Machine Learning, Natural Language Processing, Image Processing and Speech Processing. He has 28 years of teaching experience apart from 08 years of exclusive Research Experience.