Development of Question Answering System in Marathi Language

Bharat A. Shelke[1*,] **C. Namrata Mahender**[2]

1,2 Department of Computer Science & IT, Dr. B.A.M.U, Aurangabad-431004(MS),India

ABSTRACT

Data is available in large amountson the internet. People have become interested in searching their needs on the internet as it is the simplest way of gaining information on a click. Thus, the QA system is very important in retrieving the correct responses. Even "Question Answering" systems are needed in the education system. It is observed that during Covid-19 pandemic online education became the generic way of teaching learning which really sustained the education. The form of examination in online mode is mostly MCQ and many courses require subjective assessment too. Question answering is classified in two main parts: "Open domain question answering system" and "Closed domain question answering system". Our work concerns the closed domain scenario. During our study, we found that in the Indian scenario few systems are available in local languages like Gujarati, Bengali, Marathi, Malayalam, Tamil, Hindi, etc. But very little work is done in Marathi language. This paper presents QA System in Marathi Language for retrieving the answer from Marathi text corpus.

Keywords—*Question, Answer, Question Answering System, NLP*

1. Introduction

A lot of studies have been done in "Natural Language Processing" considering the rise in demand and availability of online information. The user has a lot of queries related to the QA system in order to get the right answers. The QA system ideally gathers and detects the right answer from a range of text corpuses, unlike "information retrieval systems" returned by it in the entire document. A QA system can provide answers which are a lot more focused [1,2]. There are two types of QA systems – "Open domain QA system" and "Closed domain QA system". The questions and answers are associated only with specific domains in "closed domain QA system". The first ever "open domain QA system" was "Baseball QA" which was introduced in 1961 [3]. While there is no restriction for questions and answers in "open domain" one as users can ask anything on any field to seek answers like Quora, ask.com, Google, etc. [4].

In every QA system the main thing is that questions are always asked and answers are retrieved from corpus. In a closed domain question answer system, we have to select the corpus. Corpus may be in English or any regional language. We have developed a closed domain QA system.

2. Question and Answer

2.1. What is a “Question”?

A “Question” refers to a sentence which is used to ask something from an individual, a group, a company, or search engine. A question is a sentence and a noun which is expressed or recorded in order to gather an answer [1].

2.2. How many types of Questions Asked?

Here are the common types of questions that are generally asked [4] -

Definition-type Questions – These types of questions demand answers in brief para to know the characteristics of an object. For example, when asked about an individual, there are chances that his name, birth-date, height, marital status, educational background, career details, etc. are required.

Factoid – These are basically the type of questions which need a specific fact as an answer, such as, who is the current president of India?

True/False or Yes/No questions – These types of questions are normally close-ended. They need answers in either yes or no or true or false. For example, Is Delhi really a highly polluted city?

Explanation-based – As the name suggests, such types of questions need proper explanation. For example, How did the Britishstart ruling India?

Instruction-based – These types of questions need answers on the basis of steps, instructions, or process. For example, how to bake Black Forest cake?

List – This type of question requires a list of items or anything as an answer. For example, what ingredients are required in Black Forest cake?

2.3. What is “Answer”?

The term “answer” refers to a sentence or word which is written or given as a reaction, statement, etc. to deal with a question or situation. There may be multiple expressions and words to answer a specific question. It tends to be a response to a unique aspect of an individual user.

3. Review on QA System

Indians have been inspired to work more and more on their native language with the introduction of TREC. Some of the researchers on QA systems have switched their focus from global languages like Mandarin, Japanese, and English to regional languages in India like Bengali, Tamil, Hindi, and Malayalam etc., while studying authors. In comparison to other regional languages, the works on Marathi are at a very nascent stage [2]. The “PRASHNOTTAR QA system” was developed by Sahuet et al. [4] in Hindi.

It relies on the knowledge of meaning of the question which is given and expressing the same in “query logic language”. The researchers tested the text data stored in Hindi language. They gathered the textual data in Hindi language which is gathered online. The survey was conducted by Shelke et al. [5] on “QA systems” which are made in regional languages of India. There is a lack of work in the regional language of India and there is a lack of research on Marathi. A lot of Question Answering systems which are introduced rely on “linguistic approach”. A search engine was proposed in Hindi language named “TALASH” by Vasniket et al. [6]. There are three models used to develop this framework to enhance the query on the basis of user context, lexical variance, and both techniques combined. Researchers got a good response by combining both techniques rather than using a single technique. A web application was developed by Shalini et al. [7] in Hindi. It is used to gather answers for a specific question in Hindi language. It understands the question’s meaning and parses it to gather answers. Some other models or techniques have been proposed in a vast body of research in Indian languages [8-11], but there is still a lack of research in Marathi.

4. Proposed Methodology

The present works showcase the development of our first “Question AnsweringSystem in Marathi language”.

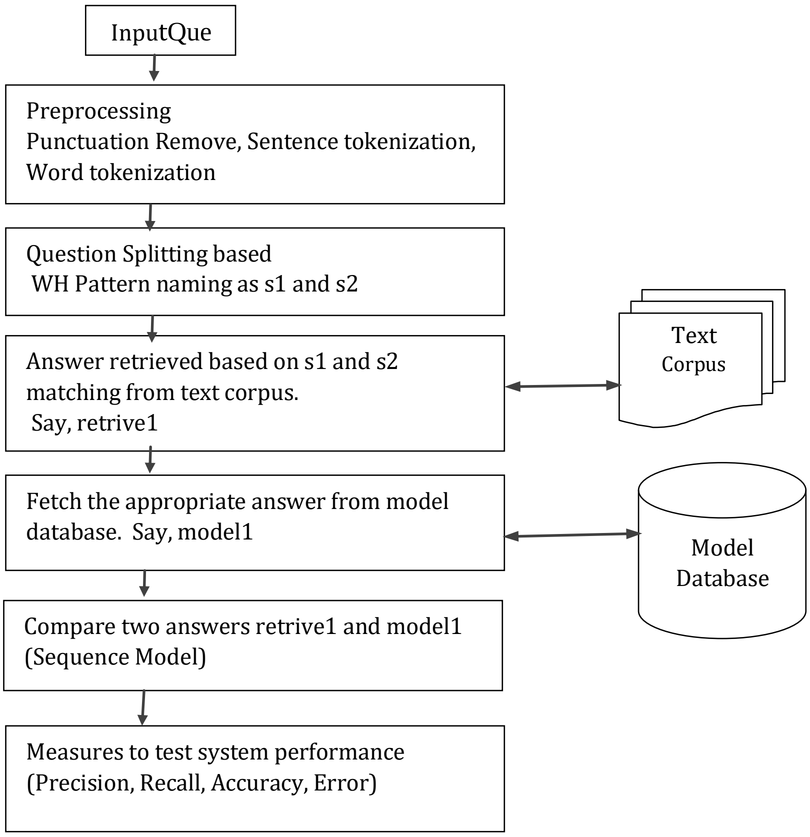

InputQue
Preprocessing
Punctuation Remove, Sentence tokenization, Word tokenization
Question Splitting based
WH Pattern naming as s1 and s2
Answer retrieved based on s1 and s2 matching from text corpus.
Say, retrive1
Text Corpus
Fetch the appropriate answer from model database. Say, model1
Model Database
Compare two answers retrive1 and model1 (Sequence Model)
Measures to test system performance (Precision, Recall, Accuracy, Error)

Figure 1 : Proposed QA System in Marathi Language

In this proposed work, we are designed a system that retrieved the answer of a question from a text corpus and also fetched the appropriate answer from a model answer database. Here we use a method of splitting a given question into two strings says s1 and s2 based on WH pattern. Every question contains the question words like केंव्हा,कोठे,किती,कायetc. There are fifty (50) question types. S1 is left part of string and s2 is right part of string based on question word. Next step is to search matches s1 and s2 in text corpus and retrieved that sentence as answer of given question named as retrive1. Our system also fetched answers from a model database named model1. To compare answers from text corpus (retrive1) and fetched answer from database (model1), sequence model is applied.

4.1. Marathi CorpusSelecton

To develop the Marathi Corpus, we are taking lessons from the 2nd, 3th and 4th classes of the Balbharati Marathi text book. The chapter summary is written. There are forty-two (42) lessons, we are taking only six (06) lessons summary for research work. We developed a Marathi text corpus in six lessons.

4.2. Database Design and Development [12]

The database is the most crucial principle in the creation of the QA system. We created our own database in order to implement this framework. Forty-two lessons were taken from the Balbharti book. We have created nine hundred one (901) questions and answers. These question answers are stored in the ‘qas_ans’ database as shown in Table1.

Table 1: Class Wise Questions and Lessons

Class	**Questions**	**Lessons**
Class – II	401	23
Class – III	212	09
Class – IV	288	10
Total	**901**	**42**

4.3. Model Question Answer Database

There are three classes, forty two (42) lessons. For model question answer database creation, we selected only two (02) lessons from each class and designed ten (10) questions on each lesson.There are six (06) lessons and sixty (60) questions are designed for model databases. Model question answers are stored in the ‘mod_ans’ database. Questions are framed and then referred to the subject experts for spelling and grammar mistakes. Some sample question answers are shown in the Table 2.

Table 2: Sample of Model Question Answer Database

Class	Model Question	Model Answer
दुसरी	केशवरावकोणत्यागावातीलशेतकरीहोते?	केशवरावभरतपूरगावातीलशेतकरीहोते.
दुसरी	केशवरावांचाव्यवसायकोणताहोता?	केशवरावांचाव्यवसायशेतीहोता.
दुसरी	केशवरावांनीमुलांनाप्रतेकीकितीरुपयेदिली?	केशवरावांनीमुलांनाप्रतेकीदोनरुपयेदिली.
दुसरी	केशवरावांनीमुलांनाकायआणायलासांगीतले?	केशवरावांनीमुलांनाघरभरूनजाईलअशीवस्तूआणायलासांगीतली.

4.4. Question Classification

We worked on fifty (50)types of question in Marathi Language such as कधी,कशाची, काय, कुठे, केंव्हाetc. These question types are shown in Table 3.

Table 3: Question Classifications

कधी	करीत	कशल्या	कशा	कोणी
कशाची	कशाचे	कश्याच्या	कशात	कोणासोबत
कशानी	कशाने	कशापासून	कशामुळे	कोणासारखे
कशाला	कशावर	कशासाठी	कशी	कोणासाठी
कसली	कसा	कसे	का	कोणाला
काय	कायम	कितवा	कितवी	कोणामध्ये
किती	कुठे	कुणी	केंव्हा	कोणाबद्दल
कोठून	कोठे	कोण	कोणता	कोणाजवळ
कोणती	कोणते	कोणत्या	कोणाकडून	कोणाच्या
कोणा	कोणाकडे	कोणाचा	कोणाची	कोणाचे

5. Evaluation Metrics and Similarity Measures

5.1 Evaluation Metrics

To check the performance and the accuracy of the QA system, we have used four measures such as Precision, Recall, Accuracy, Error%. Precision and Recall are the most commonly used measures in Natural Language Processing. In Precision correct and wrong answers were calculated, while in Recall measure missed values were calculated. Here the formulas of precision and recall are as follows:

$$Precision = \frac{Correct}{(Correct + Wrong)} \times 100 \qquad (1)$$

$$Recall = \frac{Correct}{(Correct + Missed)} \times 100 \qquad (2)$$

For Accuracy measure, it just checks total correct answers with total number of questions asked. The formula is

$$Accuracy = \frac{No.of\ Correct\ \ Answers}{Total\ No.of\ Questions\ \ Asked} \times 100 \qquad (3)$$

Error percentage measure performed on incorrect answers, this is the only measure which is performed on irrelevant answers.

$$Error\% = \frac{No.of\ Incorrect\ \ Answers}{Total\ no.of\ Questions\ \ Asked} \times 100 \qquad (4)$$

5.2 Similarity Measures

The “similarity measure” refers to the similarity of two answers or data objects to be measured by a unit. In a “text similarity”, the similarity measure refers to the distance having dimensions which show the characteristics of objects. If there is a small distance, there will be a high amount of similarity. Low level of similarity is found in the large distance [13]. The range of 0 to 1 is used to measure similarity.

5.2.1 Sequence Matcher Percentage

Sequence Matcher is a class and it is used to compare two input sequences or strings. In other words, this class is useful to use when finding similarities between two strings on the character level. Aim of sequence matcher is given two sequences and finds the length of the longest subsequence present in both of them.

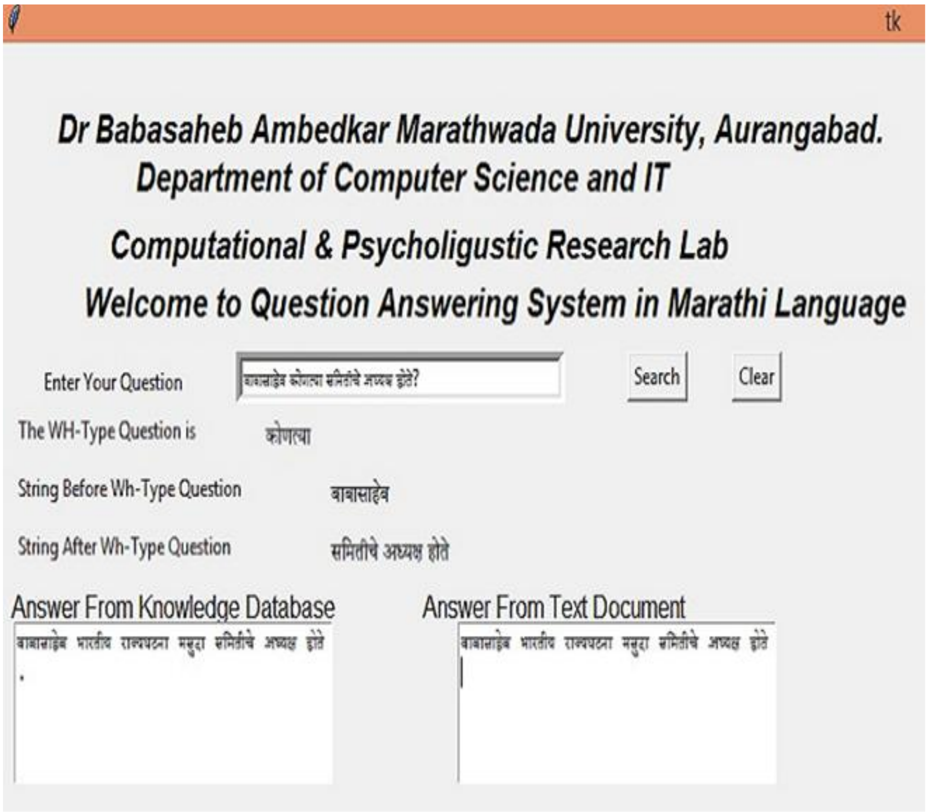
5.2.2 Levenshtein Distance [14]

The “Levenshtein Distance” metric is known to measure similarity of texts and distance between two different words. It consists of applications like “auto correction” and “text auto completion”. The word a user enters is compared to a dictionary’s words to come up with nearest matches where it makes suggestions for one of those use cases. There might be thousands of words entered in a dictionary and it may take a few ms to compare two words in an application. Basically, the “Levenshtein distance” is estimated by a matrix of “(M+1)x(N+1)” in which N and M refer to the “lengths of two words” as well as looping with the given matrix to conduct some estimations in each of those iterations.

6. Results

N-gram based type of approach is developed for retrieving answers from the Marathi corpus. We have applied a sequence model. Also, our system fetched the answer from the model question answer database. Comparison of the answers from text corpus and answers fetched from the database is done. Sequence matcher percentage similarity and Levenshtein distance similarity. System is developed using pythonAnaconda and the database is stored in MYSQL and fetched in Pandas software for processing.

In Marathi language fifty (50) types of questions are observed. We have created a database that contains nine hundred one (901) questions with answers. We validate the student answers with model answers and then perform results on the basis of evaluation metrics such as Precision, Recall, Accuracy and Error%. We are tested on two hundred (200) questions on this system. We are asked different types of questions such as: काय (71), कोठे (16), कोणत्या (12), का (11), कोण (11), कसं (08), कशामुळे (07), कशी (05) etc. We are developed Question Answering System in Marathi Language and system’s GUI interface is shown Figure2.



Dr Babasaheb Ambedkar Marathwada University, Aurangabad.
Department of Computer Science and IT

Computational & Psycholigustic Research Lab

Welcome to Question Answering System in Marathi Language

Enter Your Question: बाबासाहेब कोणत्या समितीचे अध्यक्ष होते?

The WH-Type Question is: कोणत्या

String Before Wh-Type Question: बाबासाहेब

String After Wh-Type Question: समितीचे अध्यक्ष होते

Answer From Knowledge Database: बाबासाहेब भारतीय राज्यघटना मसुदा समितीचे अध्यक्ष होते .

Answer From Text Document: बाबासाहेब भारतीय राज्यघटना मसुदा समितीचे अध्यक्ष होते

Figure – 2:GUI of Question Answering System

Table 4: Result analysis of QA System

Questions Asked	Correct Answers	Wrong Answers	Missing Answers
200	140	47	13

We observed that we asked two hundred (200) questions to the system, out of 200, correct answers 140, wrong answers 47 and missing answers were 13. Our QA system results are Precision 74.87, Recall 91.5, Accuracy 70 and Error% is 23.5.

There are two hundred (200) answers that are fetched from the database and our corpus. These answers are compared using the similarity measures. We used a sequence matcher similarity and levenshtein distance similarity. We observed that one hundred forty (140) answers having a similarity percentage is above 90. Average sequence matcher percentage is 79.68 %.

In Table 5 and 6, we used some notations such as model1 is the answer fetched from model database and retrieve1 is the answer retrieved from Marathi corpus. ‘S’ denotes the Sequence Similarity percentage between model1 and retrieve1 and ‘L’ denotes the levenshtein distance similarity between model1 and retrieve1.

Table 5: Data Samples of Correct answers

Q	Answer fetched from model database (model1)	Answer retrieved from corpus (retrieve1)	S	L
1	गाडगेबाबांच्याहातातएककाठीअसायची.	गाडगेबाबांच्याहातातएककाठीअसायची.	100.00	0
2	झाडाखालीगवतावरससेखेळतआहेत.	ससेझाडाखालीखेळतआहेत.	71.70	15
3	तळ्यातहत्तीउभाआहे.	तळ्यातहत्तीउभाआहे.	100.00	0
4	ससेहत्तीच्यापाठीवरबसलेआहेत.	लांडग्यापासूनजीववाचवण्यासाठीससेहत्तीच्यापाठीवरबसलेआहेत.	66.67	31
5	लांडग्यापासूनजीववाचवण्यासाठीससेहत्तीच्यापाठीवरबसलेआहेत.	लांडग्यापासूनजीववाचवण्यासाठीससेहत्तीच्यापाठीवरबसलेआहेत.	100.00	0

Table 6: Data Samples of Wrong answers

Q	**Answer from model database (model1)**	**Answer retrieved from corpus (retrieve1)**	**S**	**L**
1	लांडगासस्यांनापकडण्यासाठीपळतआहे.	हत्तीआपल्यासोंडेतूनलांडग्यावरपाण्याचाफवाराकरतआहे.	48.35	41
2	घाबरल्यामुळेआपलाजीववाचवण्यासाठीससेपळतआहेत.	लांडगाघाबरूनपळतआहे.	37.14	42
3	हत्तीच्यापाठीवरतीनससेबसलेआहेत.	ससेजीववाचल्यामुळेआनंदानेनाचतआहेत.	32.88	47
4	लांडगासस्यानापकडण्यासाठीपळतआहे.	हत्तीआपल्यासोंडेतूनलांडग्यावरपाण्याचाफवाराकरतआहे.	48.35	41
5	हत्तीआपल्यासोंडेतूनलांडग्यावरपाण्याचाफवाराकरतआहे.	पाखरूआब्यांवरगातआहे.	40.51	40

Observation drawn from Table-6, Q1 answer is wrong because our question is ‘लांडगाकायकरतआहे? Question Splitting basedWH pattern naming as s1 and s2. S1 is ‘लांडगा’ and s2 is ‘करतआहे’. Next step is to search s1 and s2 in Marathi corpus and system retrieved the answer as ‘हत्तीआपल्यासोंडेतूनलांडग्यावरपाण्याचाफवाराकरतआहे’. Here s2 is matched to the sentence and our system retrieves the sentence as answer. Similarity between model1 and retrieve1 is 48.35 and levenshtein distance is 41.

The a sequence matcher similarity and levenshtein distance similarity for testing the similarity percentage between answers fetched from database and answers retrieved from corpus is done. We observed that one hundred forty (140 out of 200) answers having a similarity percentage are more than ninety (90).

7. Conclusion

Question answering systems are needed in many applications and auto assessment is the most demanding and challenging task. The “QA system” is mainly aimed to gather answers instead of complete documents to respond to the questions. In this regard a lot of work can be seen in Hindi, Malayalam, Tamil, Bengali as well as English, French etc. languages but as compared to them a lot has to be done in Marathi language. The present work focuses on retrieving answers for given questions from particular text taken as corpus of Marathi lessons. The proposed QA system is a factoid type closed domain Question Answering System. Overall system results are precision 76.92, recall 88.60, accuracy 70% and error 21%. Average sequence matcher percentage is 79.68 %.

References

[1] Liddy E.D., Natural Language Processing, Encyclopedia of Library and Information Science, 2nd ed., Marcel Decker, NY, 2001.

[2] Sunil A. Khillare, Bharat A. Shelke, C. Namrata Mahender, “Comparative Study on Question Answering Systems and Techniques”, InternationalJournal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11,ISSN: 2277 128X, November 2014.

[3] Ajitkumar M. Pundge, Khillare S.A., C. Namrata Mahender, “Question Answering System, Approaches and Techniques: A Review”, International Journal of Computer Applications (0975 – 8887) Volume 141 – No.3, May 2016 .

[4] ShriyaSahu, NandkishorVasnik and Devshri Roy, “PRASHNOTTAR: A Hindi Question Answering System”, International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No April 2012.

[5] Bharat A. Shelke , Ramesh Naik , C. Namrata Mahender, “A Review on Question Answering Systems for Indian Languages”, Remarking An Analisation. P: ISSN NO.: 2394-0344 E: ISSN NO.: 2455-0817, VOL-3, ISSUE-12, (Part-2) March- 2019.

[6] NandkishorVasnik, ShriyaSahu, Devshri Roy, “TALASH: A Semanticand Context Based Optimized Hindi Search Engine”, International JournalofComputer Science, Engineering and Information Technology(IJCSEIT), Vol.2, No.3, June 2012.

[7] Shalini Stalin, Rajeev Pandey, RajuBarskar, “Web BasedApplication forHindi Question Answering System”, International Journal of Electronics and Computer ScienceEngineering ISSN- 2277-1956, 2012.

[8] Vishal Gupta, “Hindi Rule Based Stemmer for Nouns”, InternationalJournal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 1, ISSN: 2277 128X, January 2014.

[9] Vishal Gupta, “Suffix Stripping Based Verb Stemming for Hindi”,International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 1, ISSN: 2277 128X, January2014.

[10] Han L, Yu ZT, Qiu YX, Meng XY, Guo JY and Si ST.,“Research on passage retrieval using domain knowledge in Chinese question answeringsystem”, In Proceedings of IEEE International Conference on Machine Learning and Cybernetics, Vol. 5, pp. 2603-2606, 2008.

[11] Jaspreet Kaur, Vishal Gupta, “Comparative Analysis of Question Answering System in Indian Languages”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, ISSN: 2277 128X, July 2013.

[12] Bharat A. Shelke, Ramesh R. Naik, C. Namrata Mahender, “Database Creation for Marathi QA System”, SSRN Elsevier Open access journal April 29-30, 2021.

[13] Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3), 297-302.

[14] *Optimizing the Levenshtein Distance for Measuring Text Similarity*, © 2022, KDnuggets, retrieved from https://www.kdnuggets.com/2020/10/optimizing-levenshtein-distance-measuring-text-similarity.html.