# QUESTION ANSWERING SYSTEM FROM UNSTRUCTURED DOCUMENTS FOR BANGLA LANGUAGE

SAMINA TASNIA ISLAM

*(BSc Engg., MIST)*

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE IN ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

2020

## APPROVAL OF BOARD OF EXAMINERS

A thesis titled "QUESTION ANSWERING SYSTEM FROM UNSTRUCTURED DOCUMENTS FOR BANGLA LANGUAGE" submitted by Samina Tasnia Islam, ID: 1014140010, Session: 2014-2015 has been accepted as satisfactory in partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering on 20 August 2020.

## BOARD OF EXAMINERS

| | |
|---|---|
| 1. ______________________________ | Chairman |
| Dr. Mohammad Nurul Huda<br>Professor and MSCSE Director<br>Department of Computer Science and Engineering<br>United International University | (Supervisor) |
| 2. ______________________________ | Member |
| Air Commodore Md. Afzal Hossain<br>Senior Instructor, Department of Computer Science and Engineering<br>Military Institute of Science and Technology | (Co--Supervisor) |
| 3. ______________________________ | Member |
| Brig Gen Mohammad Sajjad Hossain<br>Head, Department of Computer Science and Engineering<br>Military Institute of Science and Technology | (Ex-Officio) |
| 4. ______________________________ | Member |
| Lt Col Dr. Muhammad Nazrul Islam<br>Associate Professor, Department of Computer Science and Engineering<br>Military Institute of Science and Technology | (Internal) |
| 5. ______________________________ | Member |
| Dr. Muhammad Masroor Ali<br>Professor, Department of Computer Science and Engineering<br>Bangladesh University of Engineering and Technology | (External) |

# DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

Samina Tasnia Islam
Department of Computer Science and Engineering
Military Institute of Science and Technology
20 August 2020

# ABSTRACT

This research describes the design and development of an intelligent question answering technique for Bangla language. Currently, an information retrieval (IR) system does not provide specific answer of a user's question but extracts documents that are relevant to the need of a user. In this study, the intelligent question answering system is similar to information retrieval as users submit question for finding answers from the given source documents. The objective of this research is to develop a proper information or answer retrieval system rather than document or passage retrieval which IR does actually. Here, the rule based and the machine learning based approaches are implemented just like as the current research trend. For a given question, all of the approaches in this study extract keywords, lexical and semantic features from the input to find concise information or answer from the single or multiple text articles that is relevant to the question. For time and quantity related questions, the system gives specific answer otherwise the system retrieves relevant information. In the experimentation stage, the comparison between the rule based and the decision tree based approaches are made. Besides, an empirical analysis based on the other existing question answering systems has been done.

# ACKNOWLEDGEMENTS

First and foremost I am thankful to Almighty Allah for His blessings for the successful completion of the thesis. My heartiest gratitude and deep respect will go to my supervisor Dr. Mohammad Nurul Huda, Professor and MSCSE Director, Department of Computer Science and Engineering, United International University and my co-supervisor Air Commodore Md. Afzal Hossain, Senior Instructor, Department of Computer Science and Engineering, Military Institute of Science Technology, Dhaka, Bangladesh for their supervision, affectionate guidance and proper directions whenever I sought one. Their patience detecting my flaws in thinking and writing have made this thesis a success.

Specially I am grateful to the Department of Computer Science and Engineering (CSE) of Military Institute of Science and Technology (MIST) for providing their all-out support during the thesis work.

I am also grateful to all the members of my thesis committee for their valuable opinions.

# TABLE OF CONTENTS

# LIST OF SYMBOLS

| | |
|---|---|
| **log** | : Logarithm Function |
| **P** | : Probability Function |
| $\sum$ | : Summation Function |

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATION

| | |
|---|---|
| **NLP** | : Natural Language Processing |
| **IR** | : Information Retrieval |
| **QA** | : Question Answering |
| **CBR** | : Case Base Reasoning |
| **ICASERT** | : International Conference on Advances in Science Engineering & Robitics Technology |
| **AI** | : Artificial Intelligence |
| **ML** | : Machine Learning |
| **KNN** | : K-Nearest Neighbour |
| **TF-IDF** | : Term Frequency and Inverse Domain frequency |
| **POS** | : Parts of Speech |
| **ASKMi** | : Name of Japanese Question Answering System |
| **AskMSR** | : Name of Japanese Question Answering System |
| **BFQA** | : Name of Bengali Factoid Question Answering System |
| **NER** | : Named Entity Recognition |
| **SI** | : Single Interrogative |
| **DI** | : Dual Interrogative |

# CHAPTER ONE

# INTRODUCTION

A question answering system from unstructured documents for Bangla Language requires Natural Language Processing (NLP) and is an extended part of Information Retrieval (IR) or text mining. Here, the system retrieves a relevant information or answer of a question from the given unstructured documents. The main objective of a question answering (QA) system is to retrieve a relevant information or answer which satisfies the users. Currently, an information retrieval (IR) system does not provide specific answer of a user's question but extracts documents that are relevant to the need of a user.

Bangla (widely known as Bengali) is not only spoken in Bangladesh but also spoken in some regions of India like West Bengal, Tripura, Assam. Because of the availability of a large amount of information in internet, research on information retrieval or question answering system is increasing day by day. In recent years several question answering systems have been developed for some other languages. There are a few question answering systems for Bangla, one of the low resource languages [1]. The developed systems for Bangla show a lower accuracy because of scarcity of a large-scale Bangla language text corpus and lack of investigation of proper machine learning based intelligent techniques. Therefore, an investigation of machine learning based intelligent techniques is necessary to develop an intelligent QA system for Bangla language.

## 1.1 Objectives

Main objectives of this thesis work are:

- To retrieve relevant answers from unstructured documents for Bangla language.
- To find out a specific answer based on keywords, lexical type and semantic features of the question using the decision tree.
- To implement an intelligent answer retrieval system using a supervised machine learning technique.

## 1.2 Overview of the Work

In this research, we introduce an information extraction approach that mainly focuses on rule based and decision tree based QA systems for Bangla language. Here, the system allows a user to input a question based on the given source text(s) in Bangla. The system provides a relevant information or answers from an unstructured document(s). For a given question, all of the approaches in this study extract keywords, lexical and semantic features from the input to find out concise information or answer from the single or multiple text articles that is relevant to the question. For time and quantity related questions, the system gives specific answer otherwise the system retrieves relevant information. In the experimental stage, the comparison between the rule based and the decision tree based approaches are made. This QA system is integrated with the n-gram language model, the keyword extraction process and a supervised machine learning approach. Besides, an empirical analysis based on the other existing question answering systems is done. Moreover, an analysis between Chinese QA system and Bangla QA system have been discussed along with difficulties and challenges for implementing an answer retrieval system for Bangla.

## 1.3 Thesis Scope

Scopes of the research work cover the following:

- Covers natural language processing, information retrieval, text mining, advanced artificial intelligence, machine learning research areas.
- Covers study of Bangla language resource in internet.
- Covers the status of Bangla research on question answering systems.
- Covers Bangla language model and grammar.

- Covers the comparison of QA systems based on Bangla and Chinese languages.

## 1.4 Research/Thesis Steps

The thesis is organized as follows:

***Chapter 2:*** Natural language-based articles were collected from the web and then several articles/ publications related to the question answering system were studied. Thus, Chapter 2 has discussed about literature review and related works.

***Chapter 3:*** Before starting the main implementation for Bangla, an equivalent system of Chinese Case Base Reasoning (CBR) QA System was implemented and studied. Also, an English template based answering system was studied for idea development purpose.

***Chapter 4:***There were several challenges and technical design issues for working with natural language processing for Bangla language. Various challenges and difficulties towards developing an answer retrieval system for Bangla language were identified and taken into consideration for implementing a QA system for Bangla language which have been discussed here.

***Chapter 5:*** This chapter has discussed about the design and development of rule-based question answering system for Bangla language. Training data set were prepared in this step. Data pre-processing and keyword extraction was also done in this step.

***Chapter 6:***In this chapter, incorporation of Decision Tree Classifier (C4.5) with the rule-based technique has been discussed. Mathematical calculation of the Decision Tree Classifier (C4.5) and also up gradation the training data-set for measurement unit have been discussed here broadly.

***Chapter 7:*** This section is about experimental procedure. Preparing testing data-set for question, doing empirical analysis using measurement metric of the implemented technique have been presented here.

***Chapter 8:*** A comparative study among the Rule based Bangla Question Answering System, Decision Tree based Answering System, Chinese answering methods, Template Based Answering approach and word/phrase based answering approach has been discussed in chapter 8.

***Chapter 9:*** In this chapter, conclusion, thesis contributions, limitations and future works of the thesis works have been presented.

**Publication:** A conference paper titled “Design and Development of Question Answering System in Bangla Language from Multiple Documents” has been published in ICASERT 2019, Dhaka, Bangladesh. Only a partial portion of the research works has been presented in the published article.

## 1.5 Flow chart of the research steps

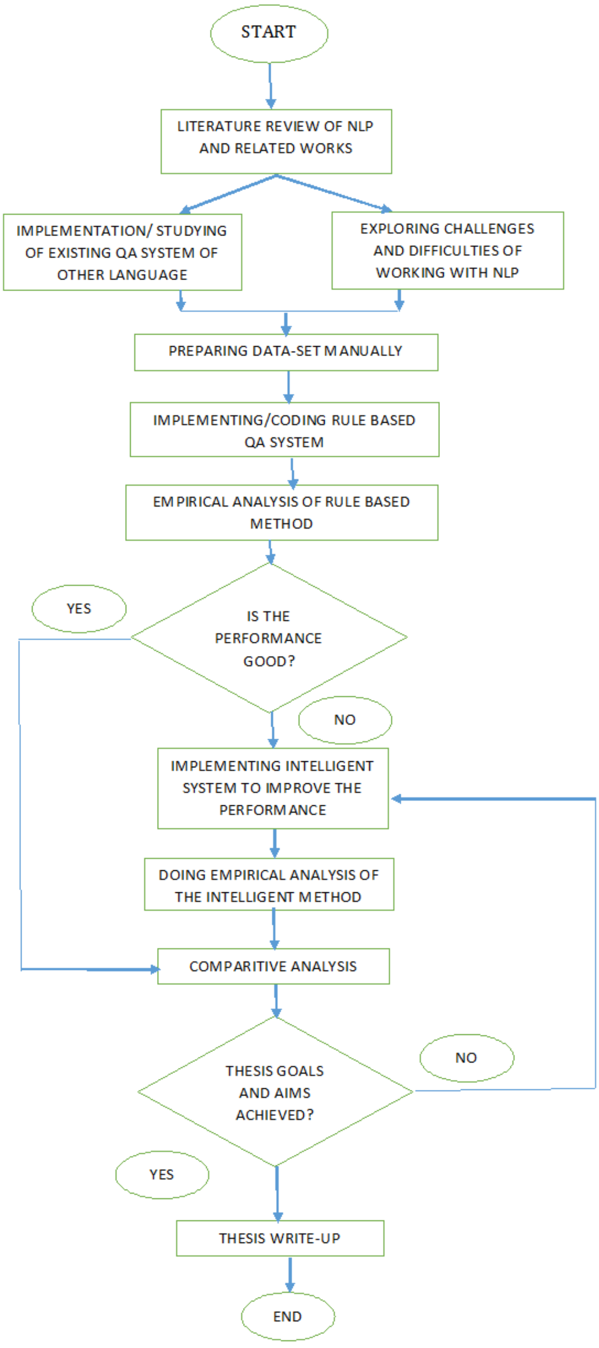

START
LITERATURE REVIEW OF NLP AND RELATED WORKS
IMPLEMENTATION/ STUDYING OF EXISTING QA SYSTEM OF OTHER LANGUAGE
EXPLORING CHALLENGES AND DIFFICULTIES OF WORKING WITH NLP
PREPARING DATA-SET MANUALLY
IMPLEMENTING/CODING RULE BASED QA SYSTEM
EMPIRICAL ANALYSIS OF RULE BASED METHOD
YES
IS THE PERFORMANCE GOOD?
NO
IMPLEMENTING INTELLIGENT SYSTEM TO IMPROVE THE PERFORMANCE
DOING EMPIRICAL ANALYSIS OF THE INTELLIGENT METHOD
COMPARITIVE ANALYSIS
NO
THESIS GOALS AND AIMS ACHIEVED?
YES
THESIS WRITE-UP
END

**Fig. 1.1.** Flow chart of the research steps

# CHAPTER TWO

# THEORETICAL BACKGROUND AND LITERATURE REVIEW

Artificial Intelligence (AI) traditionally refers to an artificial creation of human-like intelligence that can learn, reason, plan, perceive, or process natural language. Machine Learning (ML) is a part of AI and Natural Language Processing (NLP) and NLP is one of the research areas of ML or AI.

## 2.1 Preliminaries of Machine Learning

- Machine Learning (ML) is one of several AI techniques for sophisticated cognitive tasks and particularly an interesting technique because it represents a paradigm shift within AI.
- ML can solve two types of problems better than artificial intelligence techniques.These are:
  - Tasks which programmers can't describe.
  - Complex multidimensional problems that can't be solved by numerical reasoning.
- Both supervised and unsupervised approaches are used by ML. ML techniques use cross validation before using findings in real situations [2].
- Supervised learning, unsupervised learning and reinforcement learning are the three main sub-domains of machine learning [2].
  - Supervised learning requires training with labeled data which has inputs and desired outputs [2].

  - Unsupervised learning does not require labeled training data and the environment only provides input without desired target [2].
  - Reinforcement learning enables from feedback received through interactions with an external environment [2].
- Not only AI, NLP or IR research areas, many applications such as prediction of dieseases (diabetes prediction, autism spectrum disorder prediction and etc.) used ML methods [3-5].

## 2.2 Preliminaries of Natural Language Processing

- Natural Language Processing (NLP) has gained importance in the field of Machine Learning (ML) due to the critical need to understand text, with its varying structure, implied meanings, sentiments, and intent.
- NLP has removed many of the communication obstacles between machines and humans by translating machine language into human language. And so it has created opportunities for human beings to accomplish tasks which were not possible before.
- Applications of NLP include a number of fields of studies, such as natural language text processing and summarization, machine translation, speech recognition, user interfaces, multilingual and cross language information retrieval (CLIR), artificial intelligence and so on.
- Nowadays, speech recognition, document summarization, question answering, machine translation, named entity recognition, spam detection, predictive typing applications and so on are using NLP techniques.
- However, manipulation of texts for knowledge or facts extraction, for automatic abstracting and indexing or for producing content in a desired format, has been recognized as a vital research area in NLP [6].
- Moreover, a number of QA applications are now being developed that aim to give answers to users to natural language questions, as opposed to documents containing information related to the question. Such systems often use a variety of information extraction and information retrieval operations using NLP tools and methods to get

the desired answer from the given source articles [6].

## 2.3 Research Works Related with Question Answering

- **Bengali Informative Chatbot:** Bengali Informative Chat-bot used cosine similarity with TF-IDF technique to identify answers from source texts. It was a mathematical based informative chat-bot [7].
- **Doly: Bengali Chatbot:** A Bengali Chat-bot named "Doly" was presented in [8]. It could reply to a user query on behalf of a human for the education system in Bengali language. Naive Bayesian algorithm was used to generate the right answer from data.
- **Medical Related MCQ Answering System:** Medical related Automatic question answering was presented in [9]. It was implemented for English language and used USMLE data as data-set.
- **Improving Answer Extraction Using Anaphora-Cataphora Resolution:** A method for improving the answer extraction approach for single document for Bangla language using Anaphora-Cataphora Resolution was presented in [10].
- **Different Facets and Survey of Different QA System:** A survey of different developed QA methods and different facets of text based automated question answering system was presented in [11]. Another survey of QA system was presented in [12].
- **Ensemble Learning Based QA System:** Bengali question classification using ensemble learning was presented in [13].
- **Closed Domain Based Factoid QA System:** Closed domain based Bangla factoid question answering system was presented in [14]. The task was closed domain based as the knowledge base from where the answer was to be found was built based on the website which carries the information solely about Shahjalal University of Science Technology (SUST).
- **N-gram Based Passage Retrieval Engine:** D. Buscaldi, P. Rosso, J. M. G. Soriano and E. Sanchis presented a QA method based on passage retrieval technique that was focused to Question Answering. They discussed that the answer to a given question might appear in several different forms in a large collection of documents. Therefore,

it was possible to find one or more sentences that contained the answer and that also included tokens from the original question [15].

- **Word/Phrase based Answer Type Classification:** Word/phrase based answer type classification for Bengali Question Answering System was presented in [16]. The experiment was done in two setups. At first removal of the stop words then consideration of the stop words and for the later one, performance was better.
- **Comparison of ML Approaches:** A comprehensive comparison of machine learning (ML) based approaches used in Bengali question classification was presented in [17].
- **Question Classification Using Support Vector Machine:** Support vector machine based question classification was present in [18]. The main focus was classifying questions instead of developing question answering system [18].
- **Question Classification and Lexical, Syntactic and Semantic Features of Question Towards Developing a QA System:** Bengali Question classification methods were discussed in [19]. Lexical, syntactic and semantic features of Bengali question were presented in [19].
- **Natural Language Based Question Answering:** L. Hirschman, et al. had discussed about the natural language-based question answering system. They mentioned that information retrieval or extraction was related to question answering because users who provided queries, they expected to have answers of the given question from the source documents. Usually the information retrieval system returns documents instead of just answers. In such case normally, users themselves extract answers from the retrieved passages [20].
- **Automated Answer Retrieval System:** A. A. Andrenucci, et al. had reviewed as well as compared three main answer retrieval techniques: NLP based, Information Retrieval (IR) and Question Template Based QA systems. NLP based QA focused on natural language interfaces. It provided the answers by mapping text string to question string by lexical, syntactic and semantic relationship. IR based QA which was enhanced with shallow or deep NLP techniques was focused on fact extraction from a large text. A. Andrenucci, et al. had also explained that template-based QA did not

process texts. It showed that extracted information had no guarantee of appropriateness [21].

- **CBR in Chinese Language**: Case Base Reasoning (CBR), an automatic answering system was introduced in [22]. The answering system based on CBR could analyze the given question and was workable for Chinese language only. After giving an input question, the system had searched for candidate question based on the keyword of the question in historical question storage, calculated the similarity of sentences with the question and finally pre-stored answers were presented to the user [22].
- **Interactive QA in Chinese:** Zhou, et al. had taken consideration about the issues of non-user-friendliness, unintelligent and non-interactive to user question and proposed an interactive as well as intelligent QA system which was workable for Chinese language only [23].
- **Template Based Question Answering:** Advantages and disadvantages of template based question answering technique was presented in [24].
- **English Neural Network Based QA:** Yuanzhi ke and Masafumi Hagiwara had proposed an English QA system using neural network composed of five layers in [25].
- **ASKMi AskMSR in Japanese:** The first Japanese Question Answering system named "ASKMi" was presented in [26]. This technique had used Semantic Role Analysis for analyzing questions. ASKMi, was leading Japanese QA system and could do query expansion as well as document constraint generation. However, the architecture of AskMSR, another Japanese QA system, was presented in [27].
- **Factoid Question Answering:** The factoid answering system for Bengali language which was proposed in [28] named "BFQA". Based on the keywords, sentences were extracted from the document and after that sentences were ranked by the score value of the answers. Then extracted answers were validated. The developed corpus by S. Banerjee, S. K. Naskar, and S. Bandyopadhyay had the constraint that all the questions were related to a particular document only [28].
- **Survey of text QA:** P. Gupta and V. Gupta had surveyed about text question answering system as well as Indian question answering system [29-30].

## 2.4 Shortcomings of the Studied System

- Though there were some QA systems, but they were workable for foreign languages.
- Because of lexical, syntactic and semantic differences those approaches were not suitable for Bangla language.
- Already existing QA systems were domain dependent and only workable for frequently asked questions and they had used databases for answering the fixed type/ pattern oriented answers.
- Bengali Informative Chatbot extracted sentence as answers from the source texts rather generating concise answer of the question. It didn't use any machine learning algorithm like recent research trends. It could be said that the system was not able to learn by itself [7].
- Some research works were going on to develop QA systems for Bangla language, but some intelligent QA systems for Bangla language with few limitations were found. Those limitations are:
  - Doly: Bengali chatbot which used Naive Bayesian was an education based question answering system and could not answer questions from other fields [8].
  - Anaphora-Cataphora resolution based answering system was able to extract answers only from simple sentences from the source text. If sentences were complex or compound or large enough, that system got puzzled to extract answers [10].
  - Word/Phrase based answer type classification system using Stochastic Gradient Descent (SGD) classifier implemented word/sentence similarity matching based question answering technique rather than answer extraction system [16].

# CHAPTER THREE

# IDEA GENERATION AND BACKGROUND OF THE PROBLEM

## 3.1 Idea Generation

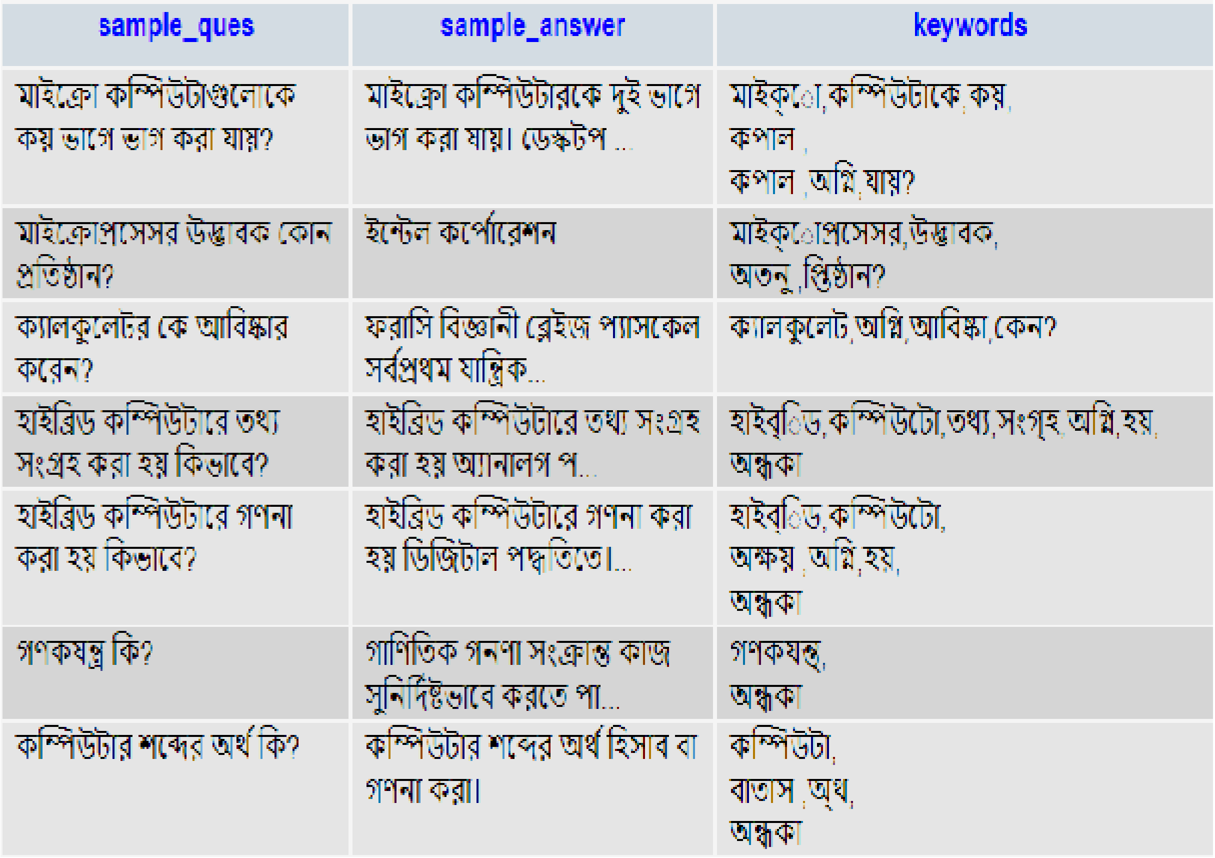
Some other existing approaches for idea generation were explored and implemented.

### 3.1.1 Chinese Case Base Reasoning QA System

This Chinese QA system known as CBR (Case Base Reasoning) was implemented [22] at first, where answers of any questions and keywords were pre-stored into a database/storage, and this domain specific application was suitable for computer related articles only.

For an input question, the QA system [22] generated the keywords from the question and then performed the matching between the generated keywords and already stored keywords. If the matching is found, the corresponding answer is retrieved; otherwise a passage is extracted by “Sentence Similarity Matching” approach and the given question was also stored for answering by a teacher in the future.

The Chinese QA system implemented for Chinese language was tested for Bangla to observe the performance. It provides poor performance because of different grammatical structure and does not work properly for other type of articles written in Chinese. Here, figure 3.1, the keywords were stemmed to find the morphological root of the word.



| sample_ques | sample_answer | keywords |
|---|---|---|
| মাইক্রো কম্পিউটাগুলোকে কয় ভাগে ভাগ করা যায়? | মাইক্রো কম্পিউটারকে দুই ভাগে ভাগ করা যায়। ডেস্কটপ ... | মাইক্ো,কম্পিউটাকে,কয়,<br>কপাল ,<br>কপাল ,অগ্নি,যায়? |
| মাইক্রোপ্রসেসর উদ্ভাবক কোন প্রতিষ্ঠান? | ইন্টেল কর্পোরেশন | মাইক্োপ্রসেসর,উদ্ভাবক,<br>অতনু ,প্তিষ্ঠান? |
| ক্যালকুলেটর কে আবিষ্কার করেন? | ফরাসি বিজ্ঞানী ব্লেইজ প্যাসকেল সর্বপ্রথম যান্ত্রিক... | কালকুলেট,অগ্নি,আবিষ্কা,কেন? |
| হাইব্রিড কম্পিউটারে তথ্য সংগ্রহ করা হয় কিভাবে? | হাইব্রিড কম্পিউটারে তথ্য সংগ্রহ করা হয় অ্যানালগ প... | হাইব্িড,কম্পিউটো,তথ্য,সংগ্হ,অগ্নি,হয়,<br>অন্ধকা |
| হাইব্রিড কম্পিউটারে গণনা করা হয় কিভাবে? | হাইব্রিড কম্পিউটারে গণনা করা হয় ডিজিটাল পদ্ধতিতে।... | হাইব্িড,কম্পিউটো,<br>অক্ষয় ,অগ্নি,হয়,<br>অন্ধকা |
| গণকযন্ত্র কি? | গাণিতিক গনণা সংক্রান্ত কাজ সুনির্দিষ্টভাবে করতে পা... | গণকযন্ত্,<br>অন্ধকা |
| কম্পিউটার শব্দের অর্থ কি? | কম্পিউটার শব্দের অর্থ হিসাব বা গণনা করা। | কম্পিউটা,<br>বাতাস ,অ্থ,<br>অন্ধকা |

**Fig. 3.1.** Sample of database for CBR

### 3.1.2 Template Based (English) QA System

A template based answering system pulls data-instances based on the question. Here, the system retrieves a question template that matches most with the given question, combines the question template with the data-instances and finally generates an answer. It is noted that this template-based approach was not domain independent, but was suitable for FAQs and structured questions. This system is appropriate where keyword or feature based typical methods retrieve an irrelevant information.

The main contribution of that system was adapting the frequently asked question answering approach to question answering technique using structured way or as a relational database system [24]. This system was appropriate only when typical questions answering was needed because conventional keyword or feature based question answering system retrieved irrelevant information.

## 3.2 Background of the Problem

QA systems are available for some languages like English, Japanese, Chinese, Hindi, Telugu, Punjabi, etc. Survey about the Indian QA systems in native language was done already [30], where the two papers were based on the Bangla language. In these two papers, one was about question classification and another one was about answering system with domain restriction.

Here, some background issues of question answering system are mentioned.

- There was no specific rule in which position the "wh" word of the question would be appeared in the sentence. So, it was a great challenge to identify the keywords from the question.
- In Bangla language it was a difficult task to map questions to answers based on the lexical, syntactic or semantic relationship between question and answer strings. The greater the answer redundancy in the document, the more possibility to retrieve an answer in a simple relation to the question.
- Automatic QA from the unstructured documents was another difficult task as there was a probability that the source text might contain only one answer to any question.
- Currently, search engines can only return a ranked list of documents, but they do not deliver the answer directly.
- Implementation of QA system based on the supervised learning was difficult because of lack of training data.
- Some existing QA systems were domain restricted also.

Therefore, a domain independent Bangla QA system for all type of documents either structured or unstructured is highly needed.

# CHAPTER FOUR

# EXPLORING CHALLENGES AND DIFFICULTIES

There were difficulties to develop an answering system for Bangla language. Technical challenges and difficulties are following:

## 4.1 Resource scarcity

- For implementing any supervised learning approaches huge data set or training set was needed but available resource was really poor for Bangla language [31].
- As there was lacking of huge amount of Bangla corpus or training data that was required for implementing supervised machine learning methods; implementing supervised learning approaches were very challenging [32].
- Bangla language is spoken in Bangladesh and also in some regions of India. For such natural language, basic language data was not available and therefore, developing resources for this natural language was still under development [30].
- The performance of a QA system depends on the knowledge sources it employed. Because of the above issues, the performance was not much better like other languages [33].
- For experimenting any QA system, a large amount of data was needed [31]. Being one of the most spoken language there were no standard question data set on the web.
- The language processing tools like POS tagger, Parser, Named Entity Recognition (NER), etc. are not publicly available for Bangla [28].

## 4.2 Difficulties in Matching Question String to Answer String

- Automatic QA from unstructured documents was challenging job since there was a probability that the main text should contain only single information or data to any user's question [27].
- Mapping of questions-answers based on lexical, syntactic and semantic relationship was another critical job [27].
- The higher the answer availability in the article, the more chance to pull an information or answer through an easy relation to the user's question; otherwise, above issues needed to be solved by NLP methods [27].

## 4.3 Issues with Keywords Extraction

- There were no specific rules in which position of the sentence the "wh" question word would be appeared. Finding keywords in this type of sentences was very important and therefore, the challenge was to identify the keyword or headword from the question [19].

## 4.4 Problem of Parts-of-Speech (POS) tagging

- POS tagging was considered as a fundamental/important element for NLP/ Information Retrieval/Text Mining applications like question answering system. In Bangla language available POS taggers which were still in research steps had different kinds of limitations [26].
- Bangla language is one of the inflected languages. Based on the suffix used in any word in Bangla language, a single word might have many forms [15].
- In Bangla language many words were modified with times and changes of actual meaning by modification made Bangla words more ambiguous as well as complex.

## 4.5 Stemming Challenges

- Stemming technique had become unsuitable for some applications because of their substantial complexities in time and space [34].
- Determining the root of the word was difficult task as Bangla is a very inflectional language and its words might have many forms. If we consider a word such as: ছেলেরা (Boys) had the derivation forms like ছেলেগুলো (Boys), ছেলেদেরকে (Boys) and etc. Similarly let consider another root word of a verb: খাওয়া (Eat) had the derivation forms like খেয়েছে (Have eaten), খেয়েছিল (ate), খাচ্ছ (eating), খাবে(will eat) and etc.
- Though there were many techniques for stemming English and later these were made workable for other languages but unfortunately none of these approaches were suitable for Bangla language because of its high inflection [35].

## 4.6 Answer Selection Problem

- Answers of any question could be appeared which were not related to each other in classical term-based searching approaches [35].
- Deciding the approximate length of definition type answer was another difficult task. There was a possibility that the answer might not appear in the retrieved passage; e.g. when retrieved passage was long enough then passage might contain some irrelevant data which were unrelated to the question-answer expected by the user. So, there was a scope of developing a better answer retrieval approaches [35] considering the issues.
- One of the vital issues in answering technique was determining which is the right answer as well as the most informative one [35].
- If the size or collections of texts was larger and more heterogeneous then extracting concise answers for the given question from this huge collection was difficult also [20].
- Traditionally information retrieval methods were considered as text or passage retrieval technique, i.e. applications which could retrieve passages or documents or

related to the user's question or query, usually didn't provide direct answers [21].

- Distinguishing the probable answer from the retrieved information of the question given by user was also challenging [21].

## 4.7 Evaluation Problem

- Deciding the criteria for evaluation of an answer retrieval system was another problem [20].

## 4.8 Domain Dependency

- Question template-based QA systems were workable only for the questions in a specific language for certain source documents. It could be said that template- based QA were domain dependent [21].

## 4.9 Rule Based QA Development Issues

- In Bangla language wh-word or interrogative could be present at any place of the question i.e., first, middle or in between and last of the question sentence. This issue made difficulties for developing a rule-based approach of question analysis.
- It was extremely challenging to generate rules for well-formed and ill-formed utterances simultaneously as people normally use ill-formed pronunciation in everyday life for communication purpose [24].

## 4.10 Interrogative Issues

- In English language there are only a few interrogatives exists but in Bangla there are many interrogatives occurs. A research presented 26 interrogatives and labeled them into three different classes. They were Unit or Single Interrogative (SI/UI), Dual Interrogative (DI) and finally Composite Interrogative or Compound Interrogative (CI) [28].

## 4.11 Un-intelligent and non-friendly QA System

- The existing QA systems for many other languages mainly offered simple mechanical answers. They were often considered as un-interactive and unintelligent. Not only these, they were also criticized for being not user-friendly with users [23].

## 4.12 Opinion Mining Problem

- Opinion or retrieved information mining from article of native or natural language is a problem of Artificial Intelligence or Machine Learning [36].
- It was extremely challenging and difficult to develop any analytical solution related with IR study [36].

The proposed system was suitable for working with more than one source documents. It eliminated the difficulties mentioned in this section such as resource scarcity on web, issues with keyword extraction, stemming challenges, answer selection problem, domain dependency, rule-based development issues, un-intelligent QA system issues and so on.

# CHAPTER FIVE

# DESIGN AND DEVELOPMENT OF RULE BASED QUESTION ANSWERING SYSTEM

The rule-based answer retrieval technique was an extended form of information retrieval based answering system. It neither required sound understanding of the language nor any sophisticated methods. In this case, normally NLP approaches were used to retrieve answers. If any question has কবে/কখন/কত, time or measurement related phrase, the answer can be retrieved through a rule. For other types of questions, descriptive or thematic answers can be retrieved.

## 5.1 Proposed Methodology of Rule Based QA System

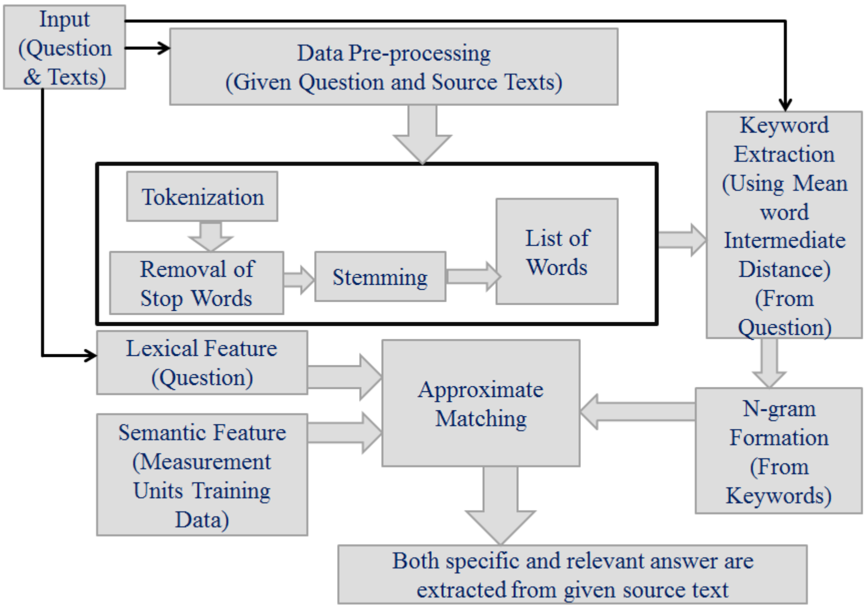

Input (Question & Texts)
Data Pre-processing (Given Question and Source Texts)
Keyword Extraction (Using Mean word Intermediate Distance) (From Question)
Tokenization
Removal of Stop Words
Stemming
List of Words
Lexical Feature (Question)
Semantic Feature (Measurement Units Training Data)
Approximate Matching
N-gram Formation (From Keywords)
Both specific and relevant answer are extracted from given source text

**Fig. 5.1.** Proposed Methodology of Rule Based QA System

Figure 5.1 shows the proposed methodology of rule based question answering system for Bangla language.

## 5.2 Data Pre-processing

In classification or information retrieval problem, feature extraction was important. It was expected that appropriate features extraction would generate the relevant information from the text. Moreover, pre-processing was necessary in NLP before feature extraction was done from documents. In pre-processing steps, each original document was considered as words of bags.

This system was designed by merging the contents of all the documents, removing stop words, stemming of question and text documents, keyword extraction from question, N-grams formation from keywords for approximate matching, retrieving n-best answers and generating specific answers by question type. Details of the steps are following:

- **Tokenization:**

  Tokenization was the task where sentences as well as text files are break down into words by white space or tab or carriage return. In this task the outcome of this process was a list of words separated by space.

- **Removing Stop Words:**

  In the proposed system necessary intelligence about conjunctions, pronouns, verbs and also inexhaustible words had to be provided for removing these stop words from the question as well as from the text documents. Approximate 65 words were manually prepared for stop words only for rule-based implementation. Though later number of stop words were increased.

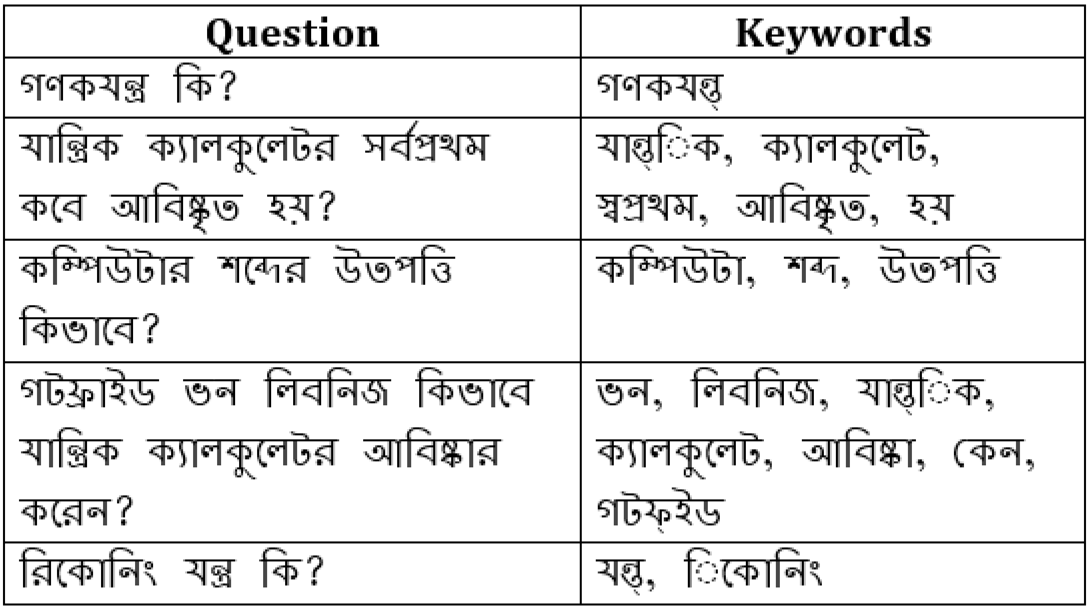
- **Stemming:**

  Intelligence about suffixes was also needed for stemming the words of the given question as well as the text document. Both the training sets of question as well as the documents were stemmed to find out the morphological stem of the words for approximate matching and retrieving information from source documents. For rule-based implementation, approximate 29 suffixes were manually annotated for stemming purpose.

## 5.3 Keyword Extraction and N-gram Formation

- **Keyword Extraction:**

  Keywords or headwords from the question had to be generated. Normally NLP related tasks use TF-IDF method for keyword extraction. After studying TF-IDF method, it was found that this method is not suitable for keyword extraction for question. It was normally workable for document categorization problems. So, the system had used a statistical idea about the mean value of the occurrence of the question-words in the given documents. Equivalent of this idea was proposed in [37]. It was not possible to consider keywords as syntactic feature because in Bangla language/grammar, there is no fixed rule for appearing the wh-word in the question.



| Question | Keywords |
|---|---|
| গণকযন্ত্র কি? | গণকযন্ত্ |
| যান্ত্রিক ক্যালকুলেটর সর্বপ্রথম কবে আবিষ্কৃত হয়? | যান্ত্িক, ক্যালকুলেট, স্বপ্রথম, আবিষ্কৃত, হয় |
| কম্পিউটার শব্দের উতপত্তি কিভাবে? | কম্পিউটা, শব্দ, উতপত্তি |
| গটফ্রাইড ভন লিবনিজ কিভাবে যান্ত্রিক ক্যালকুলেটর আবিষ্কার করেন? | ভন, লিবনিজ, যান্ত্িক, ক্যালকুলেট, আবিষ্কা, কেন, গটফ্ইড |
| রিকোনিং যন্ত্র কি? | যন্ত্, িকোনিং |

**Fig. 5.2.** Keywords extraction process

Figure 5.2 shows some generated keywords from the questions are shown. Here, the keywords are the morphological root of the word itself.

- **N-grams formation from Keywords for Approximate Matching:**

N-grams had to be formed for effective approximate matching. Keywords generated from questions would be used to form n-grams (unigram/bigram/trigram). Consequently, n-gram sequence was compared with other sequences for retrieving relevant information from the source text.

## 5.4 Answer Retrieval

- **Lexical and Semantic Features:**

  In Bangla question “wh-word” was considered a vital lexical feature. An important role was played by the end marker. If the end marker is “|”, then the given question is definition type. Lexical and semantic features helped to find out actual answer type from question. For example, if any question starts with “কে” it indicates that the user is searching an answer containing any “person names”. So, if the question starts with “কিভাবে/কেমন”, then the question indicates either adjective type or adverbial type answer. The proposed system worked for the question type of (কবে, কখন) (kəb, kəkhən) and quantity related (কত ) (kət) and so measurement unit was used as semantic feature. According to interrogative (wh-type) type and semantic feature specific answer would be retrieved.

  A document related with “Cox’s Bazar” which is one of the tourist spots of Bangladesh and another document related with “Computer” were collected from Wikipedia (Bangla) and several questions given below were set from these documents.

  – Question 1: কক্সবাজার থানা কবে প্রতিষ্ঠিত হয়? (When Cox’s Bazar Thana has been first established?)

    ∗ Answer 1: ১৮৫৪ সাল (1854 year)

    ∗ Answer 2: সাল এবং পৌসভা (Year and Municipality)

  – Question 2: কম্পিউটারে প্রথম বাংলা লেখা সম্ভব হয় কখন? (When Bangla writing was possible in Computer first?)

    ∗ Answer 1: ১৯৮৭ সাল (1987 Year)

  – Question 3: বাংলাদেশে প্রথম কম্পিউটার আসে কত সালে ? (When Computer has first come in Bangladesh?)

- ১৯৬৪ সাল (1964 Year)
- ১৯৭১ সাল (1971 Year)
- ১৯৮১ সাল (1981 Year)

In the first question the first answer is correct; the second answer here matches with the keywords of the question. In question 2, here only one answer is retrieved which is correct and in question 3 first answer is correct and other answers matches with the keywords of the question. In these answers, every word is the root of the word itself.

- **Ranking the retrieved sentences:** n-best answers were extracted from the text passages into smaller format so that it could indicate answers to the users. After that, retrieved sentences would be ranked by Textual Entailment Module (TE) [38]. And best ranked retrieved information would be considered as answer.

# CHAPTER SIX

# IMPLEMENTATION OF AN INTELLIGENT ANSWER RETRIEVAL SYSTEM

Implementation of an intelligent answer retrieval system required pre-processing of data mentioned in the previous section, algorithms of artificial intelligence and also huge data for training set. The proposed system was able to answer specific type answer for time or measurement and consequently, decision tree was implemented for such type of question-answer. For decision tree, approximate 444 stop-words and 29 suffixes were used for data pre-processing steps.

## 6.1 Proposed Methodology of Decision Tree Based QA System:

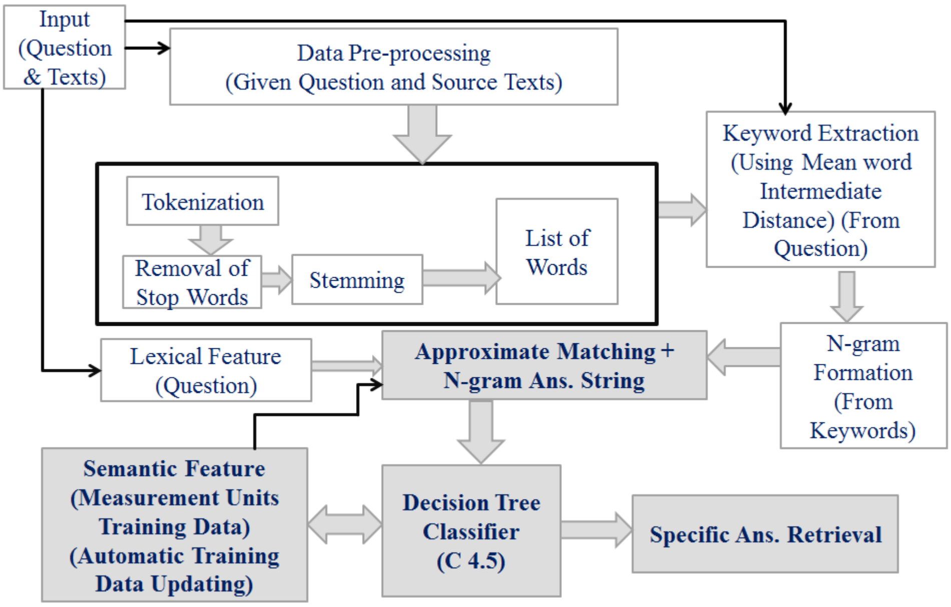

Input (Question & Texts)
Data Pre-processing (Given Question and Source Texts)
Keyword Extraction (Using Mean word Intermediate Distance) (From Question)
Tokenization
Removal of Stop Words
Stemming
List of Words
N-gram Formation (From Keywords)
Lexical Feature (Question)
Approximate Matching + N-gram Ans. String
Semantic Feature (Measurement Units Training Data) (Automatic Training Data Updating)
Decision Tree Classifier (C 4.5)
Specific Ans. Retrieval

**Fig. 6.1.** Proposed Methodology of Decision Tree Based QA System

Figure 6.1 shows the system diagram of decision tree based question answering. Selection of

Decision Tree (C 4.5) from many machine learning algorithms such as K-NN, Naive Bayes, Support Vector Machine, Neural Network, etc. was more theoretical rather practical. But working with decision tree was easier and it could work directly from the data-sets.

## 6.2 Automatic Measurement Unit/Quantifier Updating:

For specific type of answer selection, a list of quantifiers was needed. Initially 55 quantifiers training set was prepared manually by the author. Author had implemented a decision tree-based technique for updating quantifier training list automatically. Some initially prepared quantifiers are: সাল, মাস, বছর, ঘন্টা,দিন, হাজার, কোটি, টাকা, পয়সা, জানুয়ারি, ফেব্রুয়ারি, মার্চ, ডিসেম্বর ,বৈশাখ, লক্ষ , আগামীকাল, গত ,পরশু, ইত্যাদি and etc.

For this purpose, when the question type was time related or measurement related such as কবে/ কখন/ কত , N-gram was formed from the answer strings. Here, N-grams were considered as input data of the root of the tree.

## 6.3 Decision Table

**Table 6.1:** Data set for decision tree

| Numeric Data | Quantifier/ Measurement Unit | Decision |
|---|---|---|
| Exists | Exists | No need to update |
| Exists | Doesn’t exist | Need to update |
| Doesn’t exist | Exists | No Need to update |
| Doesn’t exist | Doesn’t exist | No Need to update |

The above table is about the data-set for decision tree. It was used for updating the training list of measurement unit automatically by decision tree classifier (C4.5).

## 6.4 Calculation of Decision Tree Classifier

From the decision table above, it is seen that, there are two attributes, one is Numeric Data, another one is Quantifier or Measurement Unit. Among the two attributes which one will be the root of the tree is decided by Decision Tree Classifier (C4.5). For this purpose, gain ratio of both of the attribute have to be find out.

The calculation steps are following:

$$\begin{aligned} Entropy(Decision) &= \sum - P(I).log_2 P(I) \\ &= -P(Yes).log_2 P(Yes) - P(No).log_2 P(No) \\ &= 0.81 \end{aligned} \tag{6.1}$$

Gain Ratio Calculation for the Attribute of Numeric Data:

$$\begin{aligned} Entropy(Decision|Numeric = yes) &= -P(Yes).log_2 P(Yes) - P(No).log_2 P(No) \\ &= 1 \end{aligned} \tag{6.2}$$

$$\begin{aligned} Entropy(Decision|Numeric = no) &= -P(Yes).log_2 P(Yes) - P(No).log_2 P(No) \\ &= 0 \end{aligned} \tag{6.3}$$

$$\begin{aligned} Gain(Decision|Numeric) &= Entropy(Decision) \\ &- P(Yes) * Entropy(Decision|Numeric = yes) \\ &- P(NO) * Entropy(Decision|Numeric = no) \\ &= 0.56, \end{aligned} \tag{6.4}$$

$$\begin{aligned} Split\ Info &= -P(Yes).log_2P(Yes) - P(No).log_2P(No) \\ &= 0.5 + 0.5 = 1 \end{aligned} \tag{6.5}$$

$$\begin{aligned} Gain\ Ratio(Decision|Numeric) &= Gain(Decision|Numeric)/SplitInfo \\ &= 0.56 \end{aligned} \tag{6.6}$$

Gain Ratio Calculation for the Attribute of Measurement Unit/Quantifier:

$$\begin{aligned} &Entropy(Decision|Measurement\ Units = yes) \\ &= -P(Yes).log_2P(Yes) - P(No).log_2P(No) \\ &= 0 \end{aligned} \tag{6.7}$$

$$\begin{aligned} &Entropy(Decision|Measurement\ Units = no) \\ &= -P(Yes).log_2P(Yes) - P(No).log_2P(No) \\ &= 1 \end{aligned} \tag{6.8}$$

$$\begin{aligned} &Gain(Decision|Measurement\ Units) \\ &= Entropy(Decision) \\ &- P(Yes) * Entropy(Decision|Measurement\ Units = yes) \\ &- P(NO) * Entropy(Decision|Measurement\ Units = no) \\ &= 0.311 \end{aligned} \tag{6.9}$$

$$\begin{aligned} Split\ Info &= -P(Yes).log_2P(Yes) - P(No).log_2P(No) \\ &= 0.5 + 0.5 \\ &= 1 \end{aligned} \tag{6.10}$$

$$\begin{aligned} Gain\ Ratio(Decision|\ Measurement\ Units) &= \frac{Gain(Decision|\ Measurement\ Units)}{Split\ Info} \\ &= 0.311 \end{aligned} \tag{6.11}$$

From the above calculation, it is seen that, at first, Entropy(Decision), Gain Ratio (Decision |Numeric Data) and Gain Ratio (Decision | Quantifier/ measurement unit) have been calculated to identify which will be the root node of the tree. Gain Ratio (Decision | Numeric Data) and Gain Ratio (Decision | Quantifier/ measurement unit) are 0.56 and 0.31 respectively. And so, numeric data is considered as root of the tree. Following is the diagram of updating data-set automatically.

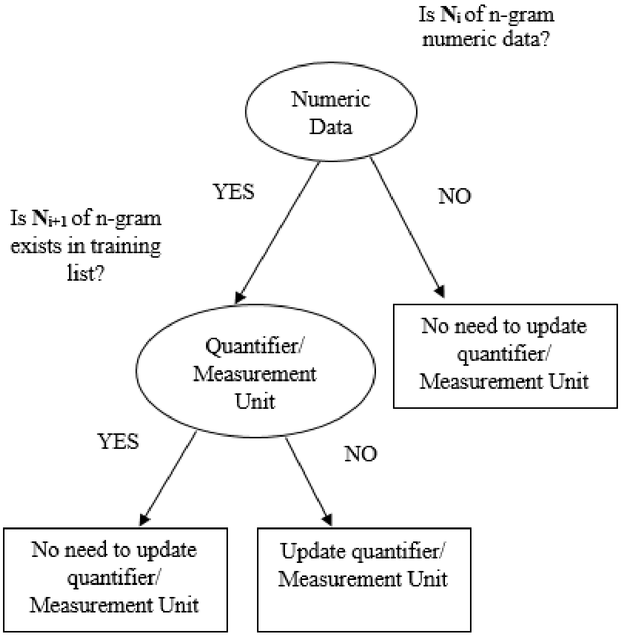

Is $N_i$ of n-gram numeric data?
Numeric Data
YES
NO
Is $N_{i+1}$ of n-gram exists in training list?
Quantifier/ Measurement Unit
No need to update quantifier/ Measurement Unit
YES
NO
No need to update quantifier/ Measurement Unit
Update quantifier/ Measurement Unit

**Fig. 6.2.** Decision tree

If any numeric data was found in the n-grams as well as n-gram followed by the numeric n-gram matched with the quantifier training list, system considered that answer was found. But if numeric data was found in the n-gram but if the following n-gram of the numeric data didn't match with the quantifier set than quantifier set was updated. If identifying correct quantifier/measurement unit is increased then, the system performances is also increased.

Document related with "Coxs Bazar", "Computer","Sundarban" and also some more documents were collected from Wikipedia (Bangla) or online news portal and approximate 500 questions were set from these documents. Given below are some of the question and answer which were retrieved using decision tree.

- Question: গণকযন্ত্র কি?
  - কম্পিউটার গণকযন্ত্র বা কম্পিউটার ইংরেজি computer কম্পিউটার হল এমন একটি যন্ত্র যা সুনির্দিষ্ট নির্দেশ অনুসরণ করে গাণিতিক গণনা সংক্রান্ত কাজ খুব দ্রুত করতে পারে
- চট্টগ্রাম শহর থেকে কক্সবাজার শহরের দূরত্ব কত ?
  - অবস্থান কক্সবাজার চট্টগ্রাম
  - অবস্থান কক্সবাজার চট্টগ্রাম শহর
  - চট্টগ্রাম শহর ১৫২
  - ১৫২ কিঃমিঃ
  - কিঃমিঃ দক্ষিণে অবস্হিত
- পশ্চিমাঞ্চলীয় সুন্দরবন অভয়ারণ্যর আয়তন কত?
  - পশ্চিমাঞ্চলীয় সুন্দরবন
  - পশ্চিমাঞ্চলীয় সুন্দরবন অভয়ারণ্যঃ
  - সুন্দরবন অভয়ারণ্যঃ ৭১,৫০২
  - ৭১,৫০২ হেক্টর
  - হেক্টর এলাকা জুড়ে
  - এলাকা জুড়ে অবস্থিত

# CHAPTER SEVEN

# EXPERIMENTAL PROCEDURES AND EVALUATION

## 7.1 Measurement Metric

Evaluation of performance and correctness of both the techniques was carried out by empirical analysis. For this purpose, the following formulas were used:

$$Precision = \frac{Relevant\ Items\ Retrieved}{Retrieved\ Item} \tag{7.1}$$

$$Recall = \frac{Relevant\ Items\ Retrieved}{Relevant\ Items} \tag{7.2}$$

$$FScore = 2 * (Precision * Recall)/(Precision + Recall) \tag{7.3}$$

## 7.2 Corpus Collection for Chinese CBR System:

- **Document Collection:** Only computer related document was selected for experimenting.
- **Question Data-set Preparation:** 50 computer related questions were used to observe the performance.
- **Stemming:** 20 suffixes were used for stemming.
- **Synonym:** Synonym list were prepared manually while experimenting Chinese CBR system for the following Bangla words:
  - Computer ( কম্পিউটার ): 32 synonyms.
  - Hardware ( হার্ডওয়্যার ): 3 synonyms.
  - Calculator ( ক্যালকুলেটর ): 13 synonyms.

## 7.3 Corpus Collection for Rule Based and Decision Tree Based QA System:

- **Stop Words:** Approximate 400 words were used as stop words for decision tree based system. For rule based system, 65 words were used as stop words.
- **Stemming:** 29 Suffixes.
- **Measurement Unit:** Initially a training set of 55 quantifiers/measurement units were prepared manually. Later this training set was updated using Decision Tree Classifier (C4.5).
- **Document Collections:** 8 documents were selected for testing purpose.
- **Question Data-set Preparing:**Approximate 500 questions were prepared manually for evaluation purpose.

## 7.4 Testing Steps

- Collecting Bangla documents from Wikipedia (Bangla) or several online news-portal.
- For testing the system, approximate 500 questions were set from the collected documents. Preparing those questions for testing purpose was lengthy and tiresome process.
- Giving input every single question into the system to extract answers.
- Among the answers of every single question, relevant items which were retrieved by the system were identified as well as relevant items in the given source texts were also identified.
- Like the above steps, retrieved items (total no. of answers) were also collected.
- Precision was calculated with the value of relevant items retrieved and retrieved items and Recall was found out by relevant items retrieved and relevant items for every single question by eq (7.1) and eq (7.2).
- After that average precision as well as average recall were calculated.
- Using eq (7.3), F-Score was calculated. For calculating F-Score, average precision and average recall were used.

## 7.5 Question Set Preparing

Among the 500 questions some questions are given below:

- কম্পিউটার শব্দের উতপত্তি কিভাবে?
- যান্ত্রিক ক্যালকুলেটর সর্বপ্রথম কবে আবিষ্কৃত হয়?
- কম্পিউটার শব্দের অর্থ কি?
- বাংলাদেশে প্রথম কম্পিউটার কবে আসে?
- অ্যাবাকাস কবে তৈরি হয়?
- জন নেপিয়ার এর অস্থি কি?
- গটফ্রাইড ভন লিবনিজ কিভাবে যান্ত্রিক ক্যালকুলেটর আবিষ্কার করেন?
- রিকোনিং যন্ত্র কি?
- গণকযন্ত্র কি?
- মাইক্রোপ্রসেসর উদ্ভাবক কোন প্রতিষ্ঠান?
- কম্পিউটারে প্রথম বাংলা লেখা সম্ভব হয় কখন?
- বাংলা ওয়ার্ডপ্রসেসিং সফটওয়্যার উদ্ভাবন করে কারা?
- মাইক্রোসফট উইন্ডোজ’ এর সঙ্গে ব্যবহারের জন্য ইন্টারফেস ‘বিজয়’ কবে উদ্ভাবিত হয়?
- কক্সবাজার নামটি কোথা থেকে এসেছে?
- ব্রিটিশ ইস্ট ইন্ডিয়া কোম্পানির সময় বাঙলার গভর্ণর কে ছিলেন?
- ক্যাপ্টেন কক্স কি সমস্যা সমাধানের চেষ্টা করেন?
- কবে কক্সবাজার থানা প্রতিষ্ঠিত হয়?

## 7.6 Precision and Recall Calculation

In experimental steps, precision, recall and F-Score were considered as measurement metric. Calculation of precision, recall and F-Score is shown in the following table:

| Question | Relevant Answers Retrieved | Retrieved Answers | Relevant Answers | Precision | Recall |
|---|---|---|---|---|---|
| কম্পিউটার শব্দের উতপত্তি কিভাবে? | 1 | 2 | 1 | 0.5 | 1 |
| যান্ত্রিক ক্যালকুলেটর সর্বপ্রথম কবে আবিষ্কৃত হয়? | 1 | 10 | 1 | 0.1 | 1 |
| কম্পিউটার শব্দের অর্থ কি? | 0 | 1 | 1 | 0 | 0 |
| ক্যালকুলেটর কে আবিষ্কার করেন? | 1 | 1 | 1 | 1 | 1 |
| গণকযন্ত্র কি? | 1 | 1 | 1 | 1 | 1 |
| মাইক্রোপ্রসেসর উদ্ভাবন করে কোন প্রতিষ্ঠান? | 1 | 1 | 1 | 1 | 1 |
| 'বিজয়' ইন্টারফেস কেন উদ্ভাবিত হয়? | 1 | 3 | 1 | 0.33 | 1 |

**Table 7.1:** Precision and Recall Calculation

In the above table, process for precision and recall calculation has been presented. For the first question in the Table, the system has pulled 2 answers. Among them only one answer is correct according to the given question, and in the source text there is one relevant answer. As a result, relevant items retrieved=1, retrieved items=2 and relevant items=1. So, according to the eq (7.1) precision (1/2=0.50) and using eq (7.2) recall (1/1=1.00) have been found. Following this similar technique, precision and recall have been calculated for every single question-answer. Then average precision and average recall have been calculated using the value of precision and recall found from every single question-answer. With the value of average precision and average recall, F-Score has been calculated using eq (7.3).

## 7.7 Empirical Analysis

After testing approximate 500 questions for rule-based answer retrieval system as well as for intelligent answering system, average precision and average recall were obtained, and the F-Score was measured using the value of average precision and also average recall.

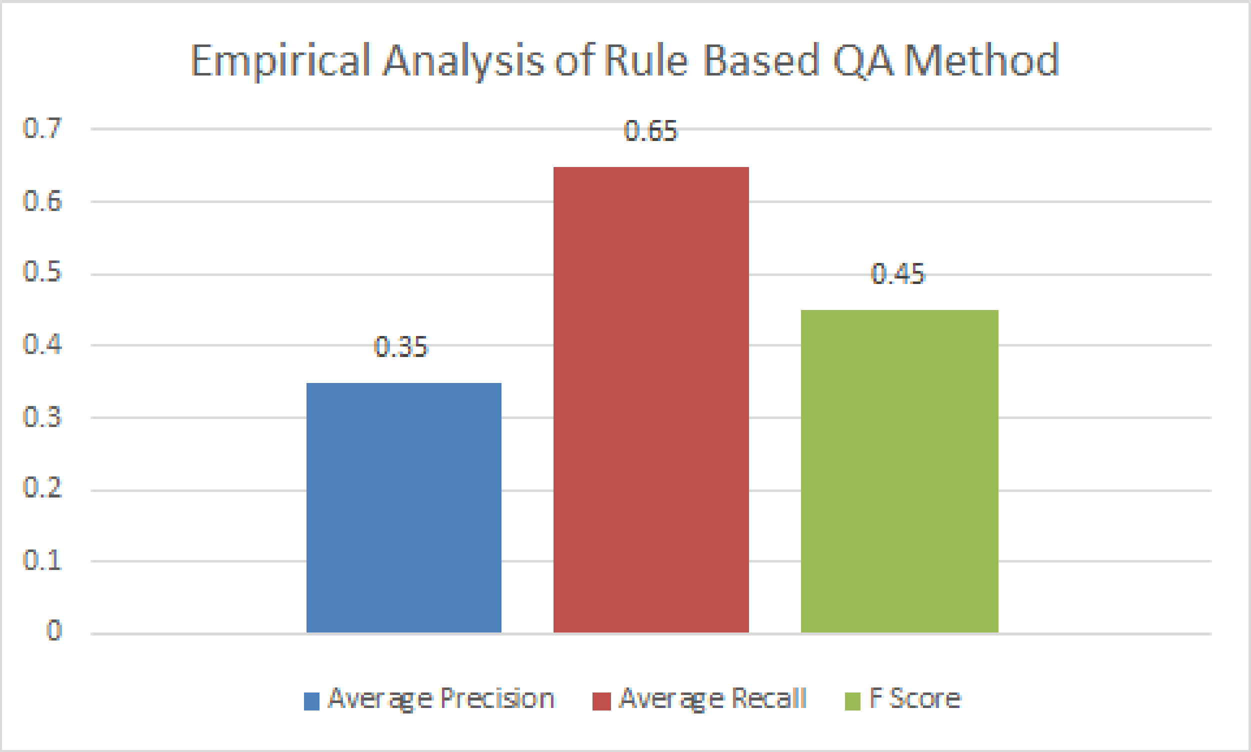


**Fig. 7.1.** Empirical analysis of rule based method

In the figure 7.1, the empirical analysis of the rule-based implementation has been shown. Here, Precision was 0.35, and recall and F-Score were 0.65 and 0.45 respectively. As the performance (F-Score) was 0.45, so after working with rule-based technique, decision tree was incorporated and tested again.

In the figure 7.2, the empirical analysis for decision tree based approach has been shown. F-score/ F-measure reveals the system performance. The higher F-score indicates that the

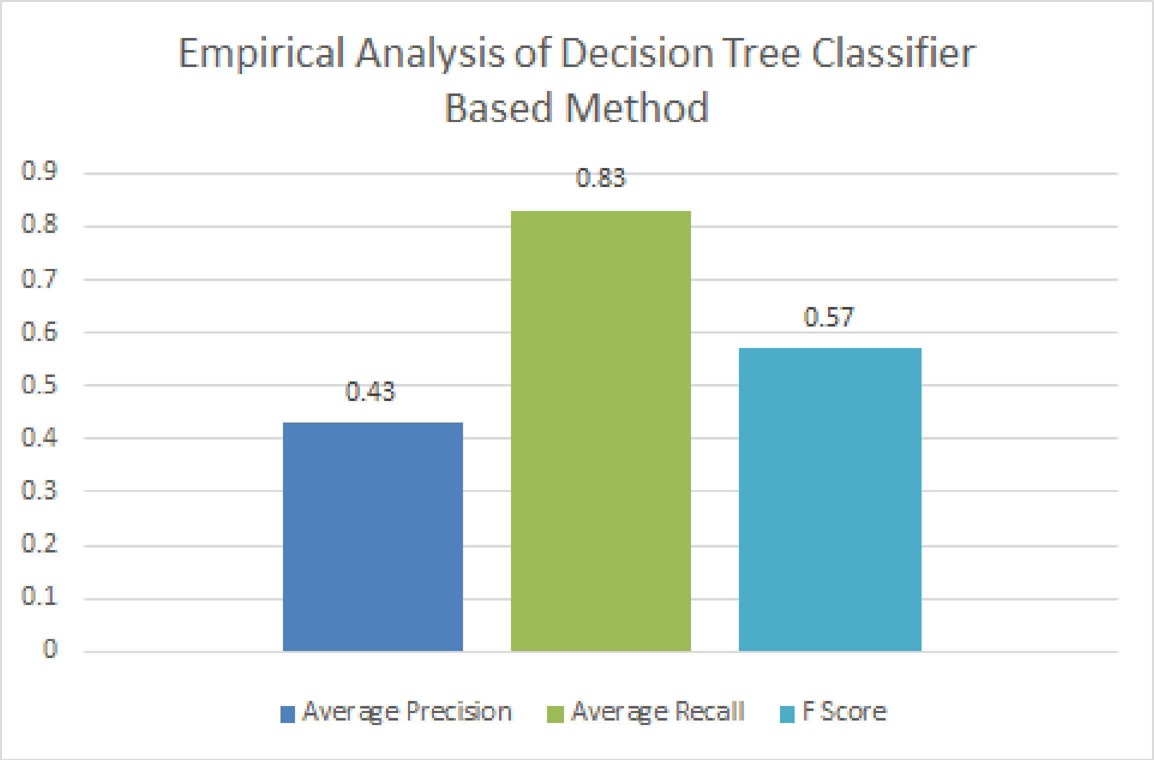


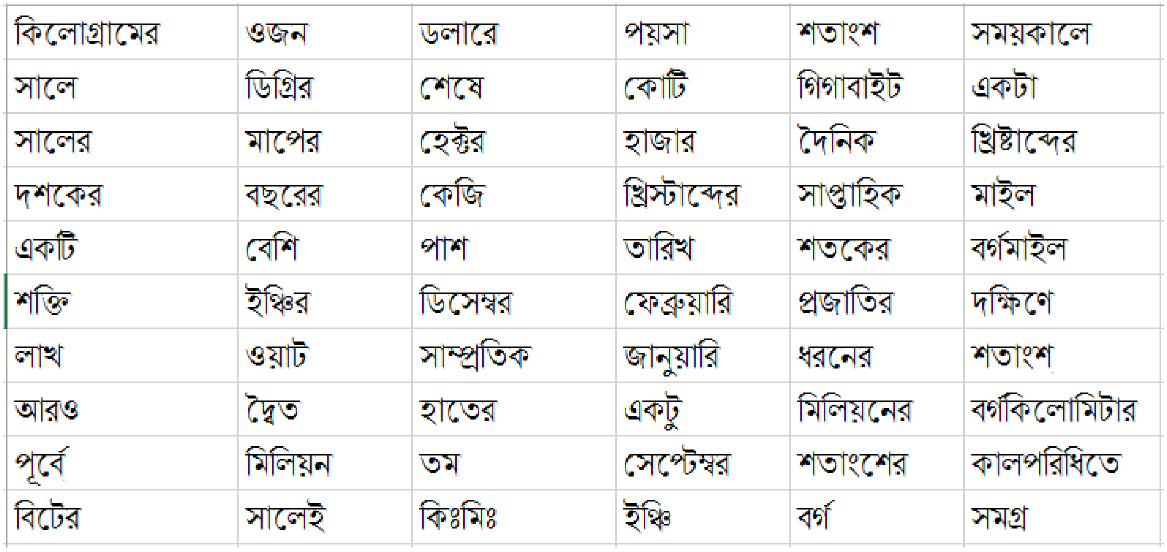
**Fig. 7.2.** Empirical analysis of decision tree classifier (C 4.5) based method

system is better. As the intelligent system updated the quantifier list automatically, it helped the system to answer more correctly for specific type question and consequently, performance is increased in such system. Here, precision was 0.43, and recall and F-Score were 0.83 and 0.57 respectively.

Precision, recall and FScore for the Chinese CBR system were 0.37, 0.48 and 0.42 respectively.

## 7.8 Result of Close Test

Intelligent answer retrieval system was able to update its training data. The system had initially 55 quantifiers, and after testing 500 questions, the system was also trained by approximate 152 quantifiers, and among them 65 were considered as correct. Following are some of the updated data by the system for specific type questions.



| | | | | | |
|---|---|---|---|---|---|
| কিলোগ্রামের | ওজন | ডলারে | পয়সা | শতাংশ | সময়কালে |
| সালে | ডিগ্রির | শেষে | কোটি | গিগাবাইট | একটা |
| সালের | মাপের | হেক্টর | হাজার | দৈনিক | খ্রিষ্টাব্দের |
| দশকের | বছরের | কেজি | খ্রিস্টাব্দের | সাপ্তাহিক | মাইল |
| একটি | বেশি | পাশ | তারিখ | শতকের | বর্গমাইল |
| শক্তি | ইঞ্চির | ডিসেম্বর | ফেব্রুয়ারি | প্রজাতির | দক্ষিণে |
| লাখ | ওয়াট | সাম্প্রতিক | জানুয়ারি | ধরনের | শতাংশ |
| আরও | দ্বৈত | হাতের | একটু | মিলিয়নের | বর্গকিলোমিটার |
| পূর্বে | মিলিয়ন | তম | সেপ্টেম্বর | শতাংশের | কালপরিধিতে |
| বিটের | সালেই | কিঃমিঃ | ইঞ্চি | বর্গ | সমগ্র |

**Fig. 7.3.** Updated data by the proposed system

## 7.9 User Interface

An interface was designed by the author of this thesis and it was helpful at the time of testing the implemented system.

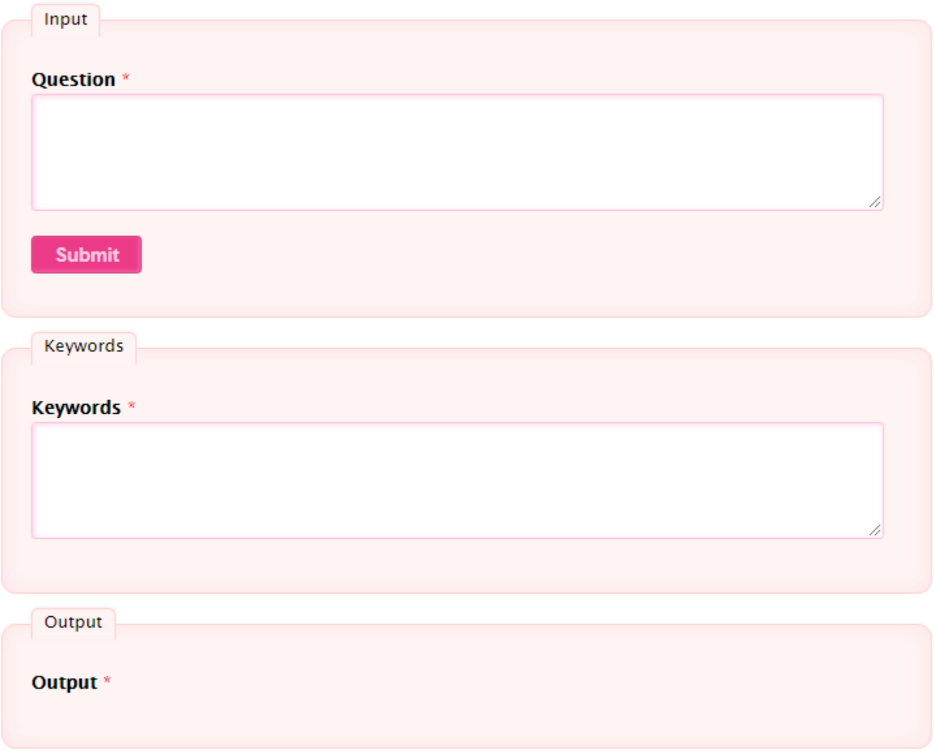


**Fig. 7.4.** User interface

In Figure 7.4, after submitting the question the system generates keywords and answers are retrieved. Both the keywords as well as n-best answers are shown in the user interface. The whole implementation has been done using PHP.

# CHAPTER EIGHT

# COMPARISON OF QA SYSTEMS

Comparison among Chinese Case Base Reasoning QA System, Rule Based Approach, Decision Tree Classifier System and Word/Phrase based question answer classification are following:

## 8.1 Comparison with CBR(Chinese) System

- Automated question answering based on CBR worked for Chinese language and the authors of this book implemented a QA system for Bangla language.
- Automated question answering based on CBR worked well for computer related courses, questions were pre-stored as training set. Whereas our system was domain independent, worked with any unstructured documents.
- In Bangla language processing data preprocessing was needed but it was not needed in Chinese question-answering system.
- As CBR of Chinese language worked for computer related courses, so after implementing an equivalent system of CBR including data preprocessing, a total of 50 computer related questions were tested. Before starting data testing, preprocessing of data was included because of inflection in Bangla language. Without data preprocessing, performance measurement was not possible for the Chinese CBR QA system.
- Precision and recall were 0.37 and 0.48 respectively, and F-Score/F-Measure was 0.42 for CBR system after including data preprocessing. Whereas for the same question set (50 questions) precision, recall and F-Score were 0.41, 0.88, 0.56 respectively for the proposed Bangla Question Answering system.

| Average Precision | Average Recall | FScore |
|---|---|---|
| 0.37 | 0.48 | 0.42 |
| 0.41 | 0.88 | 0.56 |

**Table 8.1:** Comparative Analysis between CBR and Decision Tree Classifier Methods

## 8.2 Comparison between the Rule Based and Decision Tree Based Methods

- At first a rule-based and data pre-processing-based QA system was implemented. After that supervised approach was included into the system. Following is the empirical analysis of rule-based question answering system.

| Document | Average Precision | Average Recall | F Score |
|---|---|---|---|
| 1 | 0.36 | 0.61 | 0.45 |
| 2 | 0.32 | 0.49 | 0.39 |
| 3 | 0.38 | 0.76 | 0.51 |
| 4 | 0.35 | 0.58 | 0.44 |
| 5 | 0.32 | 0.65 | 0.42 |
| 6 | 0.22 | 0.57 | 0.32 |
| 7 | 0.53 | 0.88 | 0.66 |
| 8 | 0.30 | 0.66 | 0.41 |
| Average | 0.35 | 0.65 | 0.45 |

**Table 8.2:** Average Precision, Average Recall and F-Score for Rule Based Approach

Following is the empirical analysis of decision tree-based question answering system.

| Document | Average Precision | Average Recall | F Score |
|---|---|---|---|
| 1 | 0.28 | 0.83 | 0.41 |
| 2 | 0.44 | 0.79 | 0.56 |
| 3 | 0.44 | 0.88 | 0.59 |
| 4 | 0.38 | 0.78 | 0.51 |
| 5 | 0.40 | 0.88 | 0.55 |
| 6 | 0.46 | 0.63 | 0.53 |
| 7 | 0.60 | 0.97 | 0.74 |
| 8 | 0.41 | 0.88 | 0.55 |
| Average | 0.43 | 0.83 | 0.57 |

**Table 8.3:** Average Precision, Average Recall and F-Score for Decision Tree Based System

Finally system performance increased from 0.45 to 0.57.

| Method | Average Precision | Average Recall | FScore |
|---|---|---|---|
| Rule Based | 0.35 | 0.65 | 0.45 |
| Decision Tree Based | 0.43 | 0.83 | 0.57 |

**Table 8.4:** Comparative Analysis between Rule Based and Decision Tree Classifier Based Methods

- Among the two types of implemented approaches, decision tree-based approach was able to learn by itself.
- If quantifier training data didn't contain the desired quantifier word corresponding to the question-answer then rule based approach would not retrieve correct specific answer. In such a case, decision tree based approach would update the quantifier list using Decision Tree Classifier (C4.5). Also it would retrieve answers.

## 8.3 Comparison with Template-Based Answering System

- Template-based answering system was suitable for limited, structured domain or a system which gathered question-answering knowledge step by step gradually. Though our supervised approach gathered knowledge about quantifiers but it was not restricted to domain.

## 8.4 Comparison with Word/Phrase Based Approach

- A work on identifying question answer type classification using Stochastic Gradient Descent (SGD) classifier on Bangla Question Answering system was presented in [16]. Word/phrase based question-answer type classifying method used single source document and structured data [16]. But our proposed system was designed and developed for working with multiple source documents and unstructured data.

- Words of the given question were considered as features in the word/phrase based approach. Once they removed stop words and then they considered it [16]. But according to convention of NLP, stop words have to be removed. Also because of no fixed rule of structure of Bangla question, succession of n-back to back words are not always a feature of that question. They used bigram which was formed from succession of n-back to back words of the given question and considered these as features. So, their technique was partially related to word by word sentence similarity matching between the question and the sentences of the document rather than answering a question. And so, they didn't extract keyword/ headword from question. But our proposed system extracted keywords from the question first, then these keywords were used to form n-gram for approximate matching. In our case we used trigram for answer extraction, if no answer was found for trigram then we used bigram or unigram.

- Our proposed system experimented and implemented machine learning based answer retrieval method and did empirical analysis for answer retrieval. Instead of identifying question type, time and quantity related question types were taken into consideration for specific answer retrieval using Decision Tree Classifier (C4.5). Otherwise concise answers were retrieved. As word/phrase based method experimented on question-answer type classification using Stochastic Gradient Descent (SGD) and so they did empirical analysis for classifying question answer type.

- Though empirical analysis was done on different task but their accuracy was high as they used biased data-set. They experimented on classifying question-answer type for not only factoid questions but also descriptive questions but considered very few descriptive or non-factoid questions as their descriptive questions were classified wrongly [16]. So, they considered those questions for which their accuracy would be better.

# CHAPTER NINE

# CONCLUSION AND DISCUSSION

## 9.1 Conclusion

This article has discussed question answering system from Bangla text documents. It concludes the following:

- This approach provided proper answer of a given QA system that was implemented for natural/Bangla language.
- Answer retrieval system pulled answers of different given questions instead of extracting information from system which was done by many search engines.
- Difficulties of implementing an answer retrieval system in natural language were discussed.
- This technique pulled relevant specific information or answers for measurement (time and quantity) related questions; Otherwise relevant answers were retrieved.
- Precision of rule-based technique and intelligent answer retrieval approach were 0.35 and 0.43 respectively. Recall of rule-based technique and intelligent answer retrieval approach were 0.65 and 0.83 respectively. F-Score of rule-based technique and intelligent answer retrieval approach were 0.45 and 0.57 respectively.
- There was a probability that if the documents having more than one sentence contain tokens from the given question then answering technique might pull less relevant answers since total number of retrieved answers increases. System was able to learn by itself. It was able to update its quantifier training list. It updated 152 data as quantifiers among them 65 were considered as correct.

## 9.2 Thesis Contributions

Contributions of the thesis are following:

- **Contribution to NLP Research Fields:** This thesis explored and identified all possible technical challenges and difficulties in the field of NLP for Bangla language. This would help those who are working in the field of NLP to develop any application related to Bangla language processing.
- **Contribution to Information Retrieval/Text Mining Areas:**The thesis had a contribution in text mining/ information retrieval to extract concise information (answer) using machine learning method.
- **Contribution to Implement ML approaches:**ML approaches require a huge training dataset. The thesis updated one of the training dataset automatically during question answering which might be helpful for implementing any supervised learning method.
- **Contribution to QA Field:** The thesis integrated Bangla in the QA system. So contributed to the QA field by designing and developing an intelligent QA system for Bangla language.
- **Contribution to the Bangla Corpus:** Collection of Bangla corpus was a contribution because Bangla is a low resource language and rule based approach doesn't work with noisy data whereas machine learning approaches work with noisy and clean data.
- **Contribution to Real-Life Solutions:** The experimented method could be used in any Chatbot, FAQ, Robotics and etc. Using the experimented QA system with any applications would improve the quality of the service.

## 9.3 Limitations of the Thesis

The thesis had limitations like other good research works. Limitations are given below:

- The thesis implemented only one ML technique.
- The primary limitation of the thesis was the lack of sufficient data to train the prediction model.
- The Experimental result was not compared with existing ML based QA techniques.

Rather,the experimental result was compared with rule based QA technique and the rule based technique was also part of the thesis and so this was not adequate to prove the supremacy of the proposed method.

- The thesis did not replace pronoun and synonym of any words and so, the proposed system could not generate proper answers if there were synonym or pronoun replacement issues.
- Generated n-best answer(s) were not always at top of answers sequences.

## 9.4 Future Works

The authors would like to do the following research work based on the proposed:

- Identifying other question types for Bangla language and retrieving specific answers accordingly.
- POS tagging could be implemented for better performance.
- Finding out more relevant information or answers when the retrieved answers starting with pronoun. Because when any relevant answers starting with pronoun, there was a possibility that more relevant answers might exist in different sentences of the source texts.
- More supervised learning approaches could be implemented for answering system.
- Implementation of synonym replacement so that performance could increase. For this purpose, synonym training set was needed.
- Incorporation of Adaptive learning to make the training automatically.

# REFERENCES

[1] S. Sharma, et al., “Automatic Question and Answer Generation from Bengali and English Text,” Computer Science & Telecommunications, vol. 54, no.2, April 1, 2018.

[2] J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, “A survey of machine learning for big data processing,” EURASIP Journal on Advances in Signal Processing, vol 1, pp. 67, December 1, 2016.

[3] N. S. Khan, M. H. Muaz, A. Kabir and M. N. Islam, “A Machine Learning-Based Intelligent System for Predicting Diabetes,” International Journal of Big Data and Analytics in Healthcare (IJBDAH), vol 4, no. 2, pp.1-20. July 1, 2019.

[4] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, “A machine learning approach to predict autism spectrum disorder,” in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-6, February 7, 2019.

[5] N. S. Khan, M. H. Muaz, A. Kabir and M. N. Islam, “Diabetes predicting mhealth application using machine learning,” in 2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 237-240, December 18, 2017.

[6] G.Chowdhury, “Natural language processing,” Annual Review of Information Science and Technology, 37. pp. 51-89, 2003.

[7] M. Kowsher, et al., “Bengali Informative Chatbot,” in International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), pp. 1-7, July 11, 2019.

[8] M. Kowsher, et al., “Doly: Bengali Chatbot for Bengali Education,” in the 1st Inter-

national Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1-6, May 3, 2019.

[9] L. A. Ha and V. Yaneva, “Automatic Question Answering for Medical MCQs: Can It Go Further than Information Retrieval?,” in RANLP, 2019.

[10] S. Khan, K. T. Kubra and M. M. H. Nahid, “Improving Answer Extraction For Bangali Q/A System Using Anaphora-Cataphora Resolution,” in International Conference on Innovation in Engineering and Technology (ICIET), pp. 1-6, December 27, 2018.

[11] V. Singh, “Different Facets of Text Based Automated Question Answering System,” International Journal for Research in Applied Science Engineering Technology (IJRASET), vol.6, January, 2018.

[12] E. M. N. Alkholy, M. H. Haggag and A. Aboutabl, “QUESTION ANSWERING SYSTEMS: ANALYSIS AND SURVEY,” International Journal of Computer Science Engineering Survey (IJCSES), vol.9, no. 6, December, 2018.

[13] S. Banerjee and S. Bandyopadhyay, “Ensemble Approach for Fine-Grained Question Classification in Bengali,” in 27th Pacific Asia Conference on Language, Information, and Computation, pp. 75-84. 2013.

[14] S. Sarker, S. T. A. Monisha, M. M. H. Nahid, “Bengali Question Answering System for Factoid Questions: A statistical approach,” in International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1-5, September 27-28, 2019.

[15] D. Buscaldi, P. Rosso, J. M. G. Soriano and E. Sanchis, “Answering questions with an n-gram based passage retrieval engine,” Journal of Intelligent Information Systems, vol. 34, no. 2, pp. 113-134, April 1, 2010.

[16] M. A. Islam, M. F. Kabir, K. A. Mamun, M. N. Huda, “Word/Phrase based Answer Type Classification for Bengali Question Answering System,” in 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 445-448, May 13, 2016.

[17] A. Anika, M. H. Rahman, S. Islam, A. S. M. M. Jameel, C. R. Rahman, “A Compre-

hensive Comparison of Machine Learning Based Methods Used in Bengali Question Classification," in International Conference on Signal Processing, Information, Communication Systems (SPICSCON), PP. 82-85, November 28, 2019.

[18] S. M. H. Nirab, M. K. Nayeem and M. S. Islam, "Question Classification Using Support Vector Machine with Hybrid Feature Extraction Method," in 20th International Conference of Computer and Information Technology (ICCIT), IEEE, pp. 1-6, December 22, 2017.

[19] S. Banerjee and S. Bandyopadhyay, "Bengali question classification: Towards developing qa system," in Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, pp. 25-40, December, 2012.

[20] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," natural language engineering, vol 7, no.4, pp. 275-300, December 1, 2001.

[21] A. Andrenucci and E. Sneiders, "Automated Question Answering: Review of the Main Approaches," in Third International Conference on Information Technology and Applications (ICITA'05), vol. 1, pp. 514-519, JULY 4, 2005.

[22] L. Zhenqiu, "Design of automatic question answering system base on CBR," Procedia Engineering, vol. 29, pp. 981-985, January 1, 2012.

[23] F. Zhou and B. Yang, "The Design and Implementation of an Interactive Intelligent Chinese Question Answering System," in International Conference on Intelligent Systems and Knowledge Engineering 2007, Atlantis Press, October, 2007.

[24] E. Sneiders "Automated question answering using question templates that cover the conceptual model of the database," in International Conference on Application of Natural Language to Information Systems, Springer Berlin Heidelberg, June 27, 2002.

[25] Yuanzhi ke and Masafumi Hagiwara, "An English neural network that learns texts, finds hidden knowledge, and answers questions," Journal of Artificial Intelligence and Soft Computing Research, Springer Berlin Heidelberg, vol 7, no. 4, pp. 229-242, October 1, 2017.

[26] T. Sakai, et al., "ASKMi: A Japanese question answering system based on semantic role analysis," in RIAO, pp. 215-231, April 26, 2004.

[27] E. Brill, S. Dumais and M. Banko, "An analysis of the AskMSR question-answering system," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, JULY, 2002.

[28] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "Bfqa: A bengali factoid question answering system," in International Conference on Text, Speech, and Dialogue, Springer, Cham, September 8, 2014.

[29] P. Gupta and V. Gupta, "A survey of text question answering techniques," International Journal of Computer Applications, vol. 53, no. 4, January 1, 2012.

[30] P. Gupta and V. Gupta, "A survey of existing question answering techniques for Indian languages," Journal of Emerging Technologies in Web Intelligence, vol. 6.2, pp. 165-169, May 1, 2014.

[31] A. K. Mandal and R. Sen, "Supervised learning methods for bangla web document categorization," International Journal of Artificial Intelligence Applications (IJAIA), Vol. 5, no. 5, October 8, 2014.

[32] M. T. Alam and M. M. Islam, "BARD: Bangla Article Classification Using a New Comprehensive Dataset," in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, September 21, 2018.

[33] S. M. Harabagiu, M. A. Pasca and S. J. Maiorano, "Experiments with Open-Domain Textual Question Answering," in the 18th International Conference on Computational Linguistics, vol. 1, 2000.

[34] M. Z. Islam, M. N. Uddin and M. Khan, "A light weight stemmer for Bengali and its Use in spelling Checker," in Centre for Research on Bangla Language Processing, 2007.

[35] M. R. Mahmud, et al., "A rule based bengali stemmer," in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), September 24, 2014.

[36] M. A. Karim, M. Kaykobad and M. Murshed, “Technical challenges and design issues in bangla language processing,” IGI Global, April 30, 2013.

[37] S. Siddiqi and A. Sharan, “Keyword extraction from single documents using mean word intermediate distance,” International Journal of Advanced Computer Research, vol. 6, no. 25, pp. 138, July 1, 2016.

[38] P. Pakray, P. Bhaskar, S. Banerjee, B. C. Pal, S. Bandyopadhyay, A. Gelbukh, “A Hybrid Question Answering System based on Information Retrieval,” in CLEF (Notebook Papers/Labs/Workshop), January 1, 2011.

# APPENDIX A SENTENCE SIMILARITY MATCHING CODE SEGMENT

```
if ($similarKey > $maxSimilarKey) {
    $diffSimilarKey = $similarKey - $maxSimilarKey;
    $maxSimilarKey = $similarKey;
    $ky_key = count($keywords);
    $keyLike_xy = (2 * $similarKey) / ($kx_ques + $ky_key);
    $keyLong_xy = 0;
    if ($kx_ques >= $ky_key) {
        $keyLong_xy = 1 - (($kx_ques - $ky_key) / ($kx_ques + $ky_key));
    }
    else {
        $keyLong_xy = 1 - (($ky_key - $kx_ques) / ($kx_ques + $ky_key));
    }
    $likeXY = ($lemOne * $keyLike_xy) + ($lemTwo * $keyLong_xy);
    if ($likeXY >= $thresholdDb) {
        $answerStr .= $arrResultSet[‘sample_answer’];
    }

    if ($diffSimilarKey >= 0.40) {
        $sqlStr=“SELECT * FROM question_answer_keyword_sample WHERE
        sample_ques LIKE %‘$question’%’;”
        $resultSet=mysql_query($sqlStr);
        if(!$resultSet){
            echo mysql_error($resultSet);
        }

    $sampleQues=mysql_fetch_assoc($resultSet);
        if($sampleQues[‘sample_ques’]==NULL or
        $sampleQues[‘sample_ques’]==‘Need to be updated’){
            $insertQuestionSQL = “INSERT INTO
            question_answer_keyword_sample
            (ques_type,sample_ques, sample_answer,keywords)
            VALUES (‘Need to be updated’,‘$questionInput’,
            ‘Need to be updated’,
            ‘Need to be updated’)”;
            $questionResult = mysql_query($insertQuestionSQL);
        if (!$questionResult) {
            echo mysql_error($questionResult);
            } }
        }
    }
if ($similarKeyDoc >= $maxSimilarKeyDoc) {
    $diffSimilarKey = $similarKeyDoc - $maxSimilarKeyDoc;
    $maxSimilarKeyDoc = $similarKeyDoc;
    $ky_key = count($arrDoc[$i]);
    $keyLike_xy = (2 * $similarKeyDoc) / ($kx_ques + $ky_key);
    $keyLong_xy = 0;
        if ($kx_ques >= $ky_key) {
            $keyLong_xy = 1 - (($kx_ques - $ky_key) / ($kx_ques + $ky_key));
        } else {
            $keyLong_xy = 1 - (($ky_key - $kx_ques) / ($kx_ques + $ky_key)); }
        $likeXY = ($lemOne * $keyLike_xy) + ($lemTwo * $keyLong_xy);
        if ($likeXY == $thresholdDoc || $likeXY > $thresholdDoc) {
            $answerStr.= $arrDocBefore[$i];                    }
    }
    echo $answerStr;

function openFile($fileName){
    $takeInputFromFile=fopen(“$fileName”,‘r’) or die;
    ( “Unable to open”.$fileName);
    $fileContent=NULL;
    while(!feof($takeInputFromFile)){
        $fileContent.=fgets($takeInputFromFile);
    }
    return $fileContent;
}

function removeSuffixesFromStr($stringToRemoveSuffixes, $suffixesArr){
    for ($j = 0; $j < count($suffixesArr); $j++) {
        if (preg_filter(“/$suffixesArr[$j]/”, “ ”,
        “$stringToRemoveSuffixes”, $limit = 1) != NULL) {
            $stringToRemoveSuffixes = preg_filter(“/$suffixesArr[$j]/”, “ ”,
            “$stringToRemoveSuffixes”, $limit = 1); } }                return $stringToRemoveSuffixes;
}
```

# APPENDIX B CODE SEGMENT OF KEYWORD EXTRACTION

```
function findKeyWordsByMeanCountOfWordPosition($arrOfKeyWord, $documentString,
$suffixes,$fractionArrVal,$fractionArrWord){
        $positionArr=array();
        $strAfterConjRemoved=removeConjunction($documentString);
        $strAfterPronounAndConjRemoved=removePronoun($strAfterConjRemoved);
        $afterConjProVerbRem
        =removeVerbWords (explode(" ",$strAfterPronounAndConjRemoved));
        $afterSuffixesRemovedDocArr=removeSuffixesFromArr
        (explode(" ",$afterConjProVerbRem),$suffixes);
        for($i=0;$i<count($arrOfKeyWord);$i++) {
            for ($j = 0; $j < count($afterSuffixesRemovedDocArr); $j++) {
                if ($arrOfKeyWord[$i]!= NULL &
                $afterSuffixesRemovedDocArr[$j]!= NULL) {
                if (preg_match("/$arrOfKeyWord[$i]/",
                "/$afterSuffixesRemovedDocArr[$j]/" )) {
                array_push($positionArr, $j);
            } } }
    $subPos = 0;
    $countSw=0;
    if (count($positionArr) == 1) {
    array_push($fractionArrVal, $positionArr[0]);
    array_push($fractionArrWord, $arrOfKeyWord[0]);} else {
        for ($x = 0; $x < count($positionArr); $x++) {
```

```
    $subPosTwoSet = $positionArr[$x] - $positionArr[$x + 1];
    $countSw++;
    if ($subPosTwoSet < 0) {
    $subPosTwoSet = $subPosTwoSet * (-1);
    }
$subPos += $subPosTwoSet; }
    if(count($positionArr)!=0){
        $mean = $subPos / count($positionArr);
    }
$greaterMean = 0;
$lesserMean = 0;
for ($k = 0; $k < count($positionArr); $k++) {
    if ($positionArr[$k] <= $mean) {
        $lesserMean += $positionArr[$k];
    } else if ($positionArr[$k] > $mean) {
    $greaterMean += $positionArr[$k];
    }
}
$fraction=0;
if($countSw<count($positionArr) || $lesserMean==0){
    $fraction=$countSw/count($positionArr);
}else
    if(($lesserMean + $greaterMean)>0){
        $fraction = $lesserMean/($lesserMean + $greaterMean);
    }     }
if ($arrOfKeyWord[$i]!= NULL  $fraction!= NULL  $fraction <= $mean) {
    array_push($fractionArrVal, $fraction);
    array_push($fractionArrWord, $arrOfKeyWord[$i]);
    }         }     }
```

```
        for($y=0;$y<count($fractionArrVal);$y++){
                for($z=$y+1;$z<count($fractionArrVal)/2;$z++){
                        if($fractionArrVal[$y]>=0 $fractionArrWord[$z]!=“” $fractionArrVal[$y]>
$fractionArrVal[$z]){
                        $temp=$fractionArrVal[$y];
                        $tempWord=$fractionArrWord[$y];
                        $fractionArrVal[$y]=$fractionArrVal[$z];
                        $fractionArrWord[$y]=$fractionArrWord[$z];
                        $fractionArrVal[$z]=$temp;
                        $fractionArrWord[$z]=$tempWord;
                                        }       }
        for($n=0;$n<count($fractionArrWord)-1;$n++){
                for($l=$n+1; $l<count($fractionArrWord);$l++){
                if($fractionArrWord[$n]==$fractionArrWord[$l]){
                $fractionArrWord[$l]=“”;
                }               }       }
                $strFractionWord=implode(“ ”,$fractionArrWord);
                $arrFractionWord=explode(“ ”,$strFractionWord);
                $keyWordArr=array();
                $v=0;
        if(count($arrFractionWord)>2){
                for($fv=0;$fv<count($arrFractionWord);$fv++){
                if($arrFractionWord[$fv]!=NULL){
                $keyWordArr[$v]=$arrFractionWord[$fv];
                $v++;
                }               }
                return $keyWordArr; }
        else{ return $arrFractionWord; } }
```

# APPENDIX C CODE SEGMENT OF N-GRAM FORMATION

```
foreach(range(0,count($filteredQuestionWords)) as $i) {
    if($filteredQuestionWords[$i+2]!=NULL) {
        $triGramString=$filteredQuestionWords[$i] . “ ” . $filteredQuestionWords[$i+1].
        “ ” . $filteredQuestionWords[$i+2];
        array_push($triGramOutput,$triGramString);
    }
    if($filteredQuestionWords[$i+1]!=NULL){
        $biGramString=$filteredQuestionWords[$i] . “ ” . $filteredQuestionWords[$i+1];
        array_push($biGramOutput,$biGramString);
    }
    if($filteredQuestionWords[$i]!=NULL){
        $uniGramString=$filteredQuestionWords[$i];
        array_push($uniGramOutput,$uniGramString);
    }
}
```

# APPENDIX D CODE SEGMENT OF SPECIFIC ANSWER EXTRACTION

```
for($i=0;$i<count($triGramForSpeQKeys);$i++){
        $tempArr=explode( “ ”,$triGramForSpeQKeys[$i]);
        $t=0;
        $temp1=$t+1;
        $temp2=$t+2;
        $firstTempStr=removeSuffixesFromStr($tempArr[$t],$suffixesArr);
        $sndTempStr=removeSuffixesFromStr($tempArr[$temp1],$suffixesArr);
        $thirdTempStr=removeSuffixesFromStr($tempArr[$temp2],$suffixesArr);
                if(preg_match(“/$thirdTempStr/”,“/$specificKeys/”)){
                $tempStr=$sndTempStr.“ ”.$thirdTempStr;
                array_push($specAnsArr,$tempStr);
                }else if(preg_match(“/$sndTempStr/”,“/$specificKeys/”)){
                $tempStr=$firstTempStr.“ ”.$sndTempStr;
                array_push($specAnsArr,$tempStr);
                }
        }
```

# APPENDIX E CODE SEGMENT OF TRAINING DATA UPDATING

```
    for($i=0;$i<count($triGramForSpeQKeys);$i++){
        $specificKeys=openFile( “ whSpecificQuesKey”);
        $tempArr=explode(“ ”, $triGramForSpeQKeys[$i]);
        $t=0;
        $flag=0;
        $temp1=$t+1;
        $temp2=$t+2;
        $firstTempStr=removeSuffixesFromStr($tempArr[$t],$suffixesArr);
        $sndTempStr=removeSuffixesFromStr($tempArr[$temp1],$suffixesArr);
        $thirdTempStr=removeSuffixesFromStr($tempArr[$temp2],$suffixesArr);
        if(preg_match(“/(০)/”,“/‘$firstTempStr’/”)|| preg_match(“/(১)/”,“/‘$firstTemp-
Str’/”)||
        preg_match(“/(২)/”,“/‘$firstTempStr’/”)|| preg_match(“/(৩)/”,“/‘$firstTempStr’/”)||
        preg_match(“/(৪)/”,“/‘$firstTempStr’/”)|| preg_match(“/(৫)/”,“/‘$firstTempStr’/”)||
        preg_match(“/(৬)/”,“/‘$firstTempStr’/”)|| preg_match(“/(৭)/”,“/‘$firstTempStr’/”
        )|| preg_match(“/(৮)/”,“/‘$firstTempStr’/”)|| preg_match(“/(৯)/”,“/‘$firstTemp-
Str’/”))            {
        $tempStr = $tempArr[$t]. “ ” . $tempArr[$temp1];
        array_push($specAnsArr, $tempStr);
        if (!preg_match(“}/$tempArr[$temp1]/”, “}/$specificKeys/”) &&
        !preg_match(“/(০)/”,“/‘$tempArr[$temp1]’/”)&&
        !preg_match(“}/(১)/”,“/‘$tempArr[$temp1]’/”)&&
        !preg_match(“/(২)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৩)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৪)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৫)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৬)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৭)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৮)/”,“/‘$tempArr[$temp1]’/”)&&
!preg_match(“/(৯)/”,“/‘$tempArr[$temp1]’/”)
) {
$myFile = fopen(“whSpecificQuesKey”, “a”) or die(“Unable to open file!”);
fwrite($myFile, “” . $tempArr[$temp1]);
$specificKeys.=$tempArr[$temp1];
fclose($myFile);


}
}else if(preg_match(“/(০)/”,“/‘$sndTempStr’/”)|| preg_match(“/(১)/”,“/‘$sndTempStr’/”)||
preg_match(“/(২)/”, “/‘$sndTempStr’/”)|| preg_match(“/(৩)/”,“/‘$sndTempStr’/”)||
preg_match(“/(৪)/”,“/‘$sndTempStr’/”)|| preg_match(“/(৫)/”,“/‘$sndTempStr’/”)||
preg_match(“/(৬)/”,“/‘$sndTempStr’/”)|| preg_match(“/(৭)/”,“/‘$sndTempStr’/”)||
preg_match(“/(৮)/”,“/‘$sndTempStr’/”)|| preg_match(“/(৯)/”,“/‘$sndTempStr’/”))
{ $tempStr=$tempArr[$temp1].“ ” .$tempArr[$temp2];
array_push($specAnsArr,$tempStr);
if(!preg_match(“/$tempArr[$temp2]/”,“/$specificKeys/”) &&
!preg_match(“/(০)/”,“/‘$tempArr[$temp2]’/”)&&
!preg_match(“/(১)/”,“/‘$tempArr[$temp2]’/”)&&
!preg_match(“/(২)/”,“/‘$tempArr[$temp2]’/”)&&
!preg_match(“/(৩)/”,“/‘$tempArr[$temp2]’/”)&&
!preg_match(“/(৪)/”,“/‘$tempArr[$temp2]’/”)&&
!preg_match(“/(৫)/”,“/‘$tempArr[$temp2]’/”)&&
        !preg_match("/(৬)/","/'$tempArr[$temp2]'/")&&
        !preg_match("/(৭)/","/'$tempArr[$temp2]'/")&&
        !preg_match("/(৮)/","/'$tempArr[$temp2]'/")&&
        !preg_match("/(৯)/","/'$tempArr[$temp2]'/")
        ){
        $myFile = fopen("whSpecificQuesKey", "a") or die("Unable to open file!");
        fwrite($myFile, "". $tempArr[$temp2]);
        $specificKeys.=$tempArr[$temp2];
        fclose($myFile);
        }
    }else if(preg_match("/$sndTempStr/","/$specificKeys/")) {
        $tempStr = $tempArr[$t]. " " . $tempArr[$temp1];
        array_push($specAnsArr, $tempStr);
    }else if(preg_match("/$thirdTempStr/","/$specificKeys/")){
        $tempStr=$tempArr[$temp1]." ".$tempArr[$temp2];
        array_push($specAnsArr,$tempStr);
    }else {
        $tempStr=$tempArr[$t]." ".$tempArr[$temp1]." ".$tempArr[$temp2];
        array_push($specAnsArr,$tempStr);
    }
}
```