ARITHMETIC MATHEMATICAL PROBLEM SOLVING SYSYTEM IN SINHALA

De Silva K.C.E

(IT12143214)

Dissertation submitted in partial fulfillment of the requirements for the B.Sc. Special Honors Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2015

Declaration

“I declare that this is my own work and this dissertation1 does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the B.Sc. Special (Hons) degree in IT Dissertation under my supervision.

Signature of the supervisor: Date:

Abstract

Question answering (QA) concerns itself with the development of systems that can automatically and accurately answer questions posed in natural language, and draws upon fields such as information retrieval (IR), natural language processing (NLP) and machine learning. The earliest systems which can be described as QA systems, which parsed questions about and attempted to find the answer in a structured database. The proposed system we present an approach for automatically learning to solve arithmetic word question. Most of the students are scared, if not horrified, of math word problems. In general, they are thought of as difficult. By good problems, I mean multi-step problems that advance in difficulty over the grades, and foster children's logical thinking. The proposed system construct and solve a system of linear equations, while simultaneously recovering an alignment of the variables and numbers in these equations to the problem text. In this project we are going to apply some of the Natural Language Processing (NLP) techniques to analyze the Sinhala language to gain a better understanding of the language in an NLP perspective and as a step towards developing more complex tools for solving mathematical problems. The system consists of four major functions as Key word identification, mathematical problem identification, mathematical operation identification, and the neural network. And the system has being developed using techniques such as Natural Language Processing (NLP) and Data Classification. Mainly in keyword identification, and mathematical problem identification NLP plays a vital role. Inside the neural network the positive and negative sentences of a mathematical problem are identified. By contrast, in the neural network user don't tell the computer how to solve the problem. Instead, it learns from observational data, figuring out its own solution to the problem at hand, identifying the question type and according to that build an equation to solve the problem.

Acknowledgements

I would like to express profound gratitude to Prof : koliya pulasinghe, our lecturer in charge of the course Comprehensive Design and Analysis of Project (CDAP), for providing us with required guidance and support in the successful realization of our project. I would also like to express gratitude to our supervisor, Mr. Yasas Mallawaarachchi for his invaluable support, encouragement, supervision and useful suggestions throughout this project work.

Then would like to thank to our co-supervisor of this project, Mr. Anupiya Nugaliyadde for the valuable guidance and advice. He inspired us greatly to work in this project. His willingness to motivate us contributed tremendously to our project. This study would not have been such a success without his commonsense knowledge and perceptiveness. I owe him lots of gratitude for showing me this way of research. Besides, we would like to thank the authority of Sri Lanka Institute of Information Technology (SLIIT) for providing us with a good Environment and facilities to complete this project.

My deepest gratitude goes to my mother and my father for the unconditional support given and without their support, this would have been impossible. It’s their unselfish care that makes my dream come true. Last but not least I thank all who supported me to make this work a success.

March, 17, 2016 De Silva K.C.E

Table of Contents

List of table

List of figures

CHAPTER 01

1 Introduction

1.1 Project Background

In recent years, Question Answering (QA) system has been researched extensively and automatic question answering has become an interesting research field and resulted in a visible improvement in its performance. Especially, during the last decade, a number of automatic QA systems have emerged. The technology of QA relates to a lot of aspects of NLP (Natural Language Processing), such as Information Retrieval (IR), Information Extraction (IE), Automatic Summarization, Conversation Interface etc. Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. Question Answering System that can be used in automatic training and system malfunctions diagnostics.

Natural language processing which deals with understanding of languages is sub division of Artificial Intelligence. Question Answering is a classic NLP application. It has practical applications in various domains. Most of the available system's Task that a question answering system realizes is given a question and collection of documents, finds the exact answer for the question. Developing systems that are able to answer natural language questions automatically has been a long-standing research goal. Building systems that enable users to access knowledge resources in a natural way (by asking questions) requires insights from a variety of disciplines, including, Artificial Intelligence, Information Retrieval, Information Extraction, Natural Language Processing.

Currently the technology has reached to an extent that even grade one students are familiar with e-learning. Children tend to practice self-learning via such learning tools besides of what they are being taught by teachers. Natural Language Processing is a highly advanced area which shows a rapid amelioration. Under that, Question Answering systems play a vital role.

For question answering systems that use databases to find an answer, the main challenge is to transform a natural language question into a database query. Well-known database-oriented question answering system are BASEBALL, which answers questions about results, locations, and dates of baseball games. The other type of question answering systems are text-based systems. Textual question answering systems do not require their knowledge bases to be in a particular format, instead they aim to find an answer to a question by analyzing documents in plain-text format, such as newspaper/newswire articles, manuals.. Textual question answering systems match the question with text units, e.g., phrases or sentences, in the document collection, and within those units, identify the element the question is asking for.

Proposed question answering system use to solve arithmetic problems. Arithmetic is a branch of mathematics that deals with properties of the counting (and also whole) numbers and fractions and the basic operations applied to these numbers. As a matter of fact, as a noun in the above sense, the word is used quite seldom.

Given a natural language question posed by a user, the first step is to analyze the question itself and identify key word and question analysis component may include a morpho-syntactic analysis of the question. The question is also classified to determine what it is asking for, i.e., whether it is asking for "how much ", "how many"…., etc.

Mathematical problem are normally attractive because the text is relatively straightforward. The system identifies the relevant variables and the values by analyzing each of the sentences in the problem. Then it maps the information into an equation that represents the problem, and after that generates the answer.

Although there are many advanced related systems for English language, there is no any system for Sinhala language. This research will be a great opportunity to Sinhalese who are expecting this type of system in their native language. The Sinhala language may lag behind some of the languages in terms of work carried out in the field of computational linguistics. It however, lacks any linguistic resources such as large corpora for Sinhala, Efforts have been made at research level to experiment with processing the language for different computational purposes but the degrees of success of such efforts have been prevent by the poor computational linguistic infrastructure available.

1.2 Literature Review

Our main emphasize on the research is to enable a proper mechanism to solve arithmetic problems in Sinhala language. In order to come up with these ideas we tried to understand the concepts and the previously built similar systems (all of them are in English language) There are especial research based on Question Answering on IQ questions [13]. It was found that the task of a question answering system is finding the exact answer for a given question by finding out through relevant documentation on internet. Other than that none is developed in Sinhala language.

The Sinhala language is unique to Sri Lanka that evolved independently over many centuries. Proposed system provides ideal solutions for the problem such as, analyze the Sinhala language,

When identifying mathematical operation defined in the problem firstly we need to generate a meaning with the words [5] [6] [21].It however, lacks any linguistic resources such as large corpora, lexica and other NLP tools for Sinhala. Efforts have been made at research level to experiment with processing the language for various computational purposes.

What are the technologies we have to use to increase accuracy level of understanding Sinhala language of the system. The proposed system use WinPython-64bit-3.4.3.7 platform and tested for developed own Sinhala corpus Sinhala language.

Mylanguages.org [21] is a website which provides user the facility to learn nearly 100 languages including Sinhala language. Sinhala alphabet, adjectives, adverbs, numbers, nouns, articles, pronouns, plural, feminine, verbs, prepositions, negation, questions, vocabulary, phrases, reading, Sinhala keyboard, quizzes are used as learning tools to enhance the Sinhala language skills of the users.

One of the most memorable systems was BASEBALL. Although, capable of answering rather complex questions, BASEBALL was, not surprisingly, restricted to questions about baseball facts, and most question answering systems were for a long time restricted to front-ends to structured databases. However, question answering track there has been great progress in open domain question answering. These systems use unrestricted text as a primary source of knowledge.

We found many articles, reports, documents and prototypes of other language analyzing and translation systems on the World-Wide Web and accumulate much more information؛on various Machine Translation methods and the basic concept behind Natural Language Processing.

Stanford lecture series for natural language processing is an archive of audio and video recording of lectures and supplemental materials and posts is very helpful to get some knowledge about question answering. Within these videos and lecture series they have widely discussed about the question answering concepts and so on.

Apart from the above we had a brief information gathering on Sinhala part of speech (POS) tagging research groups in Sri Lanka universities and higher education centers[5] [6] . Some research groups are focused to develop Stochastic Part of Speech Tagger for Sinhala. Mostly these taggers based on Hidden Markov Models (HMMs) .

There's an approach for automatically learning to solve arithmetic problems [2] [11]. Using algorithm constructs a systems of equations while aligning their variables and numbers to problem text .whole solution specifies the questions' answers .When they define a two-step process to map word problems to equations. First, a template is selected to define the overall structure of the equation system. Next, the template is instantiated with numbers and nouns from the text. During inference they consider these two steps jointly.

Mathematical problem are normally attractive because the text is relatively straightforward. According to the paper, the system identifies the relevant variables and the values by analyzing each of the sentences in the problem. Then it maps the information into an equation that represents the problem, and after that generates the answer.

The proposed system is most similar to the 'ARIS' [11]. Which is based on variable / value analyzing, keyword extraction etc. ARIS only relies on learning verb categories which alleviates the need for equation templates for arithmetic problems In English language. But proposed system distinguishes from this system by using all above methods in Sinhala language. In this approach question word to be classified may be a noun, verb, number or adjective, etc. Each kind of question possesses its special classification problems. In this system, we used text classification techniques, and through that understanding, improve the tools that are available for text classification and training the data which we used for keyword identification. For this purpose we used naïve bayes classifier .Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute or word belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

There are so many text books published by the Sri Lankan government in order to improve the Sinhala knowledge of the Sri Lankan school students. For example the Sinhala text book published by Education Publication Department- Isurupaya in the year of 2010 June and Sinhala & Sinhala Literature text books for grade 10 and 11 students were the reference materials we used.

1.3 Research problem

Problem solving has been and will be a necessary skill not only in Mathematics but in everyday living. The students’ ability in solving word problems depends on how students translate phrases into mathematical symbols. The first is to understand the problem so that we could see clearly the given tasked, the second is devising a plan, third is to carry out the plan and the fourth is to look back at the completed solution. The existing arithmetic problem solving systems are not directly able to solve problems in Sinhala language. So we, a group of students of SLIIT are trying to offer a software solution, which may help to clear this hurdle of creating an effective arithmetic problem solving system in Sinhala language.

Proposed mathematical problem solving system that is capable of understanding a detailed sentence mathematical problem in Sinhala language. Main research problem is have to come up with a solution to understand the Sinhala language of the system and to analyze it and to build questions in the correct Sinhala grammar format and the other thing is providing the question to the user will be done by displaying it on the screen.

Since each natural language is built on its own building blocks and structures, most of the time two languages may not be able to handle in the similar manner. Regardless some Indian languages may have common features with Sinhala, they are not identical. According to the survey results we have got as well as the facts on the internet published by screen readers reviews, and they say that the best software they use to display Sinhala words on the screen is Win python. Which is able to read and print Sinhala words clearly as English.

The challenge is that not all, we should make the system to read the Sinhala language and also to understand the answer given by the user.

Another major problem is identify which problem solving methods students choose to solve a problem. Determine if students solve math problems using addition, subtraction, multiplication, and division consistently and whether students transfer these skills to other mathematical situations and solutions. Performing addition is one of the simplest numerical tasks. But for arithmetic problem solving it is different.

Mainly we are focused on people who are able to read, write and understand any reading materials which have written in Sinhala language. When they deal with our system they can get a clear idea about how the Sinhala sentence(s) which is given by the user.

Student should understand the each words and that word sense is very important for problem solving and prefer mathematics that does not focus on problem solving. Subtraction, multiplication, and division same as the addition, students was used to investigate how students solve word problems and how they determine which mathematical approach to use to solve a problem. It was discovered that many of the students read and re-read a question before they try to find an answer.

At one extreme we have sets of 'problems' which are all about practicing a technique. In the classroom this typically involves the teacher introducing a task and illustrating the technique, and then the children do lots more 'problems' on the same theme so that they master the technique .Problem solving is interpreted as working through a series of related and predictable questions in order to acquire a particular skill. These are the major issues the team gathered during the survey which awaits solutions.

1.4 Research objectives

1.4.1 Main Objective

Develop a system that is capable of solving arithmetic problems. System must be able to take an arithmetic mathematical word problem as a text, understand the question and solve the problem. Answer should be presented to the student as a simulation so that the child can understand it more easily.

1.4.2 Specific objectives

1) Maintain better interaction with user and the system.
2) Provide an attractive user interface that suits for students in age of 7 – 10.
    i. Providing a user friendly interface for users. Interfaces should be very user friendly so that students can easily get their work done through the system without wasting time on struggling how to operate the system.
3) Understand correctly the given mathematical problem by the system.
    i. System should be capable of correctly understanding the mathematical problem given as the input.
4) Perform keyword identification
    i. First of all system must split the mathematical problem into smaller entities. Out of those entities most relevant words are extracted as key words.
5) Identify mathematical question
6) Determining the meaning of the question and what they ask
7) Identify correctly the relevant mathematical operation for a given problem
    i. System is required to be capable of understanding whether an addition or subtraction should be performed.
8) Make the final output more understandable for the user .Create correct and comprehensive simulations for each problem.
    i. System must provide the student not only the final answer, but also the way that the problem was solves. The simulations must be easy to understand and attractive.
9) Create a neural network to achieve the accuracy.

CHAPTER 02

2 Methodology

Methodology is a framework or set of practices, procedures and rules, which are used to structure, plan and control the process of developing an information system. Architecture of proposed system

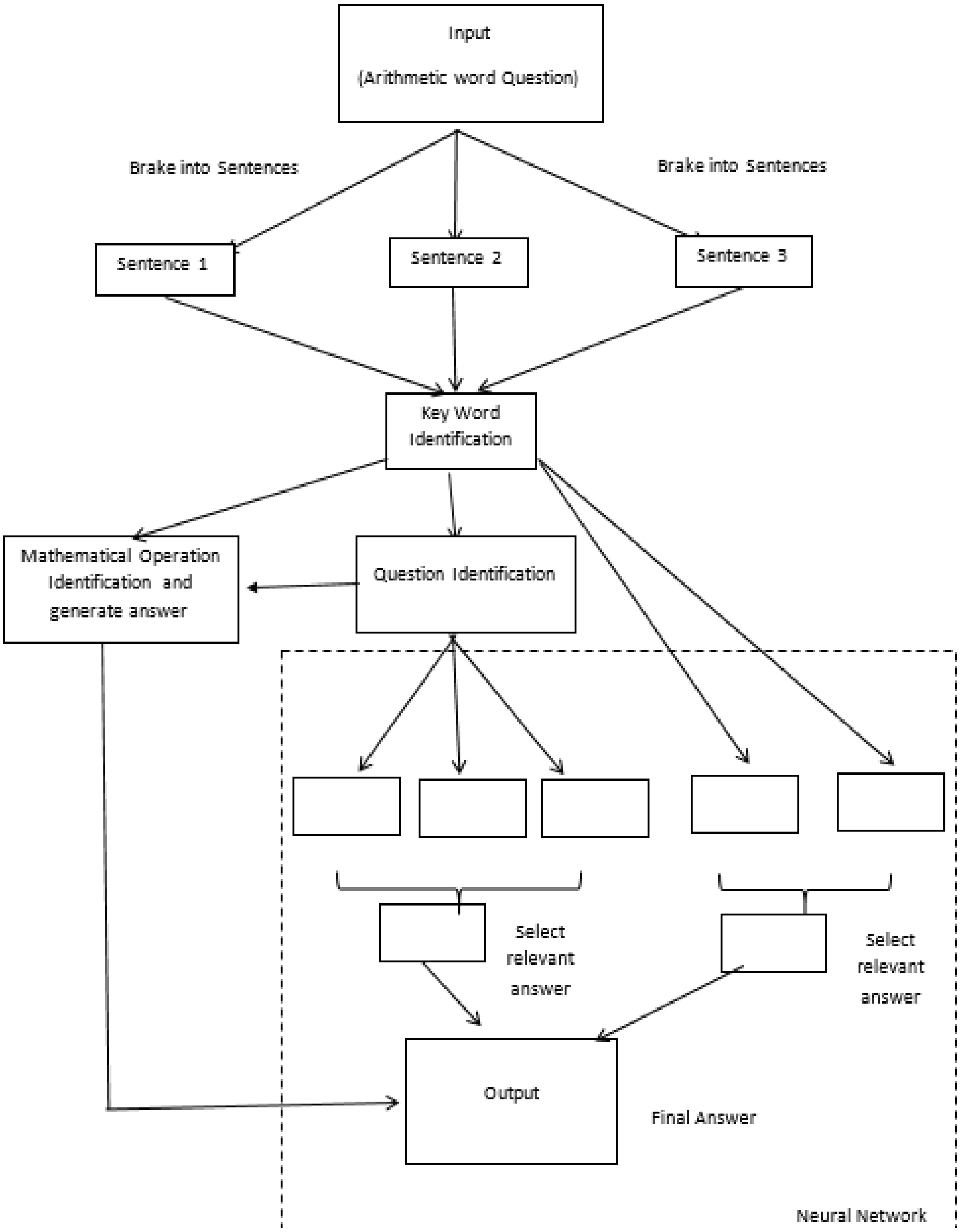

Figure 1 System diagram

Arithmetic problem solving system is the process or set of procedures to be followed to develop a platform to analyze and understand arithmetic problem in Sinhala language. Following diagram shows the set of steps to be followed in this system

This section describes the system in narrative form using non-technical terms. “Arithmetic problem solving “through neural network by using natural language techniques. This will be develop to learn how to solve mathematical problems. Under this topic it is described the development process, the methods and techniques that will be used to develop this system.

From the literature survey which was conducted to identified that most of the QA systems are designed only for the information retrieval. The functionality provided by this system it is intended to guide the user to work efficiently and reduce the time consumption to solve arithmetic problems.

The system is a combination of four components.

1. Key word identification
2. Mathematical Question identification
3. Identifying Mathematical Operation and generating an equation
4. Neural Network

2.1 Key word identification

There were some important language analysis work has done for Sinhala language, and created a Tag set and a corpus of one million words. Which was an important step that gives a considerable effect to perform NLP research on Sinhala language, For an example most frequent words in Sinhala language are function words, which belong to closed class category. 11 out of 20 words belong to prepositions which all are function words, 6 frequent verbs are also within the list. It is also observed that tagging ambiguity exists among high frequent words, though they are function words. There are two words among top 20 words, “පත්” and “යුතු” that are not properly classified into respective parts of speech categories. B progress of computational linguistic analysis on Sinhala language is far behind than other languages. According to our knowledge, there is no well-known automated POS tagging system available for Sinhala language. But in our research area we mainly focus to solve arithmetic problem in Sinhala. So we do not need much that details or more analysis about Sinhala language but we need simple idea about the language for allocating a tag to each word.

2.1.1 Part of speech tagging

The process of assigning one of the parts of the speech to the given word is called Part Of Speech tagging Parts of speech can be divided into two broad super categories: closed class types and open class types. There are four major open classes that occur in the languages of the world: nouns, verbs, adjectives, and adverbs. All four of these, although not every language does.

Proposed system select out nouns or other important words from a question. One of the main requirements of this research part was having a collection of data set which has the capability of handling data for tagging purpose and training purpose. We referred related books and Sinhala corpus to collect information about Sinhala words, as well as most of the details collected form Sinhala text books. A part from the above we had a brief information gathering on Sinhala tagger developing research group in Sri Lanka universities and higher education centers.

Typical Language Analysis Process

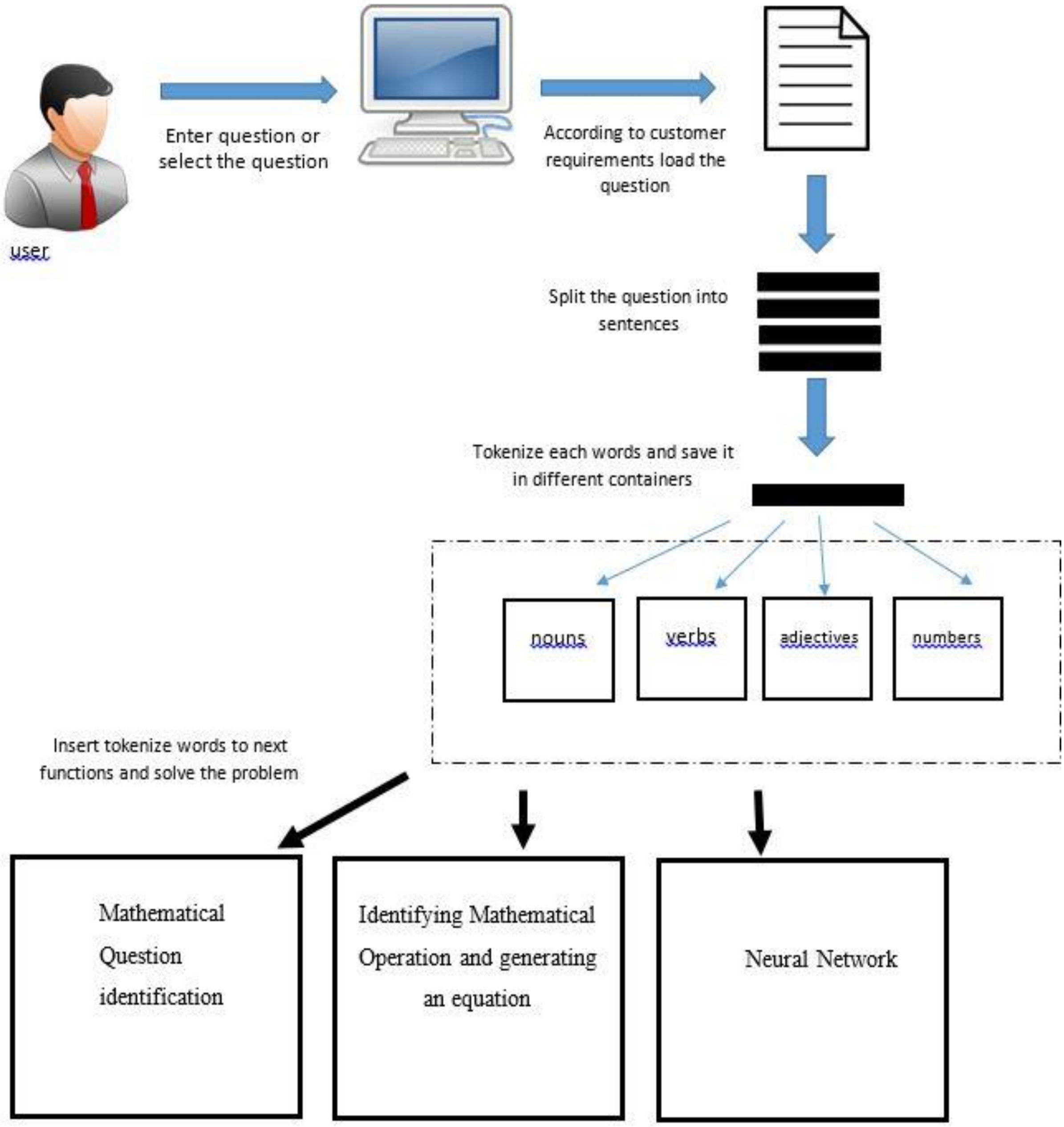
user
Enter question or select the question
According to customer requirements load the question
Split the question into sentences
Tokenize each words and save it in different containers
nouns
verbs
adjectives
numbers
Insert tokenize words to next functions and solve the problem
Mathematical Question identification
Identifying Mathematical Operation and generating an equation
Neural Network

Figure 2 Typical Language Analysis Process

Once the mathematical problem is inserted to the system, question would be divided to sentences by full stop and question mark. As a step breaking the sentence into words by separating by blank. Words would be act as the input parameters for the next functions and identify the word belongs to which category. As final step output would be generated stating that each of the word belongs to which tag.

Two kinds of ambiguous situations that arise mainly in Sinhala are handled by the model designed in the work .Known Word Ambiguity Words often belong to two or more syntactic categories (can have two or more POS Tags). When a word occurs in a given math problem, the category to which it belongs becomes clear even though the word may belong to several categories. Since it has occurred in several contexts in the training data the model knows the possible tag. Word Ambiguity When a word which has never occurred in the training phase of the model is seen, it has to depend more on the word and it will predicate that word can belongs to particular tag.

Most of the available classifications have been made with the intention of teaching/learning the language. Since this kind of classification is not suitable for an arithmetic problem solving system. So proposed system developed a more comprehensive classification for the purpose of defining POS tags.

Following table shows the set of tags used in proposed system.

Category	**Tag**
Noun	NN
Verb	VB
Pronoun	PRO
Prepositions	PP
Adjectives	ADJ
Numbers	CD
Stop words	SW
Question words	QW

Table 1 Tag set

Given a math word problem system grounds the word into tags.

For an example

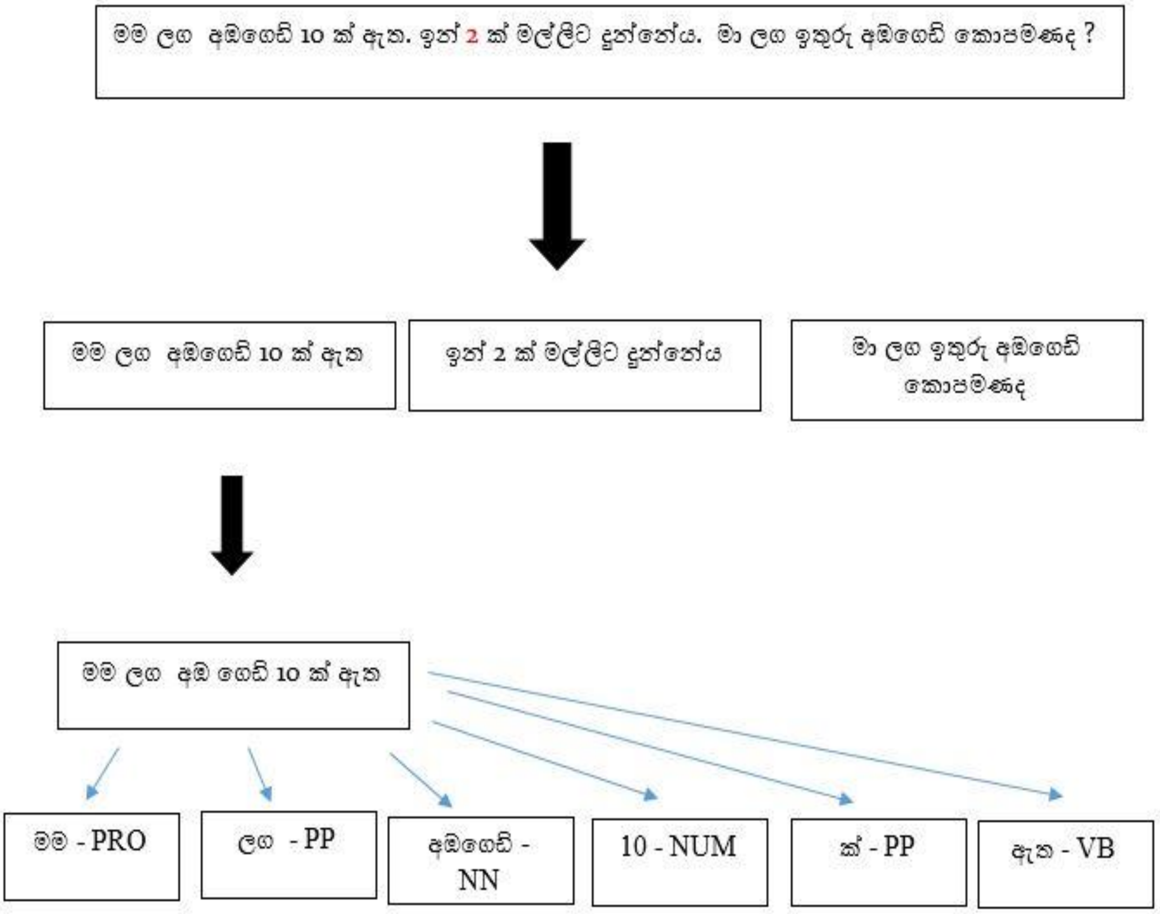

Figure 3 Example for tagging sentence

In this approach sentence word to be classified may be a noun, verb, number or adjective, etc. Each kind of question possesses its special classification problems. In this system, we used text classification techniques, and through that understanding, improve the tools that are available for text classification and training the data which we used for keyword identification. For this purpose we used naïve base classifier.

2.1.2 Naïve base classifier and training process

In general all of Machine Learning Algorithms need to be trained for supervised learning tasks like classification, prediction etc. By training it means to train them on particular inputs so that later on we may test them for unknown inputs (which they have never seen before) for which they may classify or predict etc. (in case of supervised learning) based on their learning. This is what most of the Machine Learning techniques like Neural Networks, naïve base.

As indicated, the objects (Sinhala base word) can be classified as either අම්මා_NN or අම්මගේ_NN. Our task is to classify new cases as they arrive, i.e. decide to which tag they belong, based on the currently existing objects. Since there are first letters belongs to same category it is reasonable to believe that a new case (අම්මාට, අම්මගෙන්, අම්මලා) have membership to same tag _NN. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of objects, and often used to predict outcomes before they actually happen.

A corpus is the most import resource of the tagging process because they are based on data-driven approaches since the training phase is completely based on the information contained in the corpus. Since a corpus was not available for Sinhala language and for solving arithmetic problems, relatively small one was built for this project. Which contain more than 5000 words and was selected from the text books, math papers, articles which was freely available on the World Wide Web and schools. Even though the corpus is relatively small, it has been possible to avoid the imbalance that could have arisen in the corpus because words are needed from same domain. In order to provide information about the distribution of the POS tags in the corpus, the texts in the corpus should be tagged. This tagging was done manually.

When we are doing manual tagging comparing Sinhala nouns, The Sinhala noun has four types of inflections namely Number, person, gender and determination. (English determinations are used as a separate word, a boy the boy etc.) Sinhala verb is inflectionally richer than the English verb. Sinhala verb has more than 36 inflection forms for the two voices (active and passive) and person number word inflections. Also Sinhala has 4 moods namely Indicative, Operative, Imperative and conditional. According to these ways we identify and allocate each tags manually for remaining words.

Key word identification experimentally evaluate a method and build a dataset of arithmetic word problems. We test our method on the accuracy of solving arithmetic word problems and identifying each tag categories in sentences. The accuracy of the output of system varies with the user input. Therefore through following methods the accuracy level of the system can be increased.

i. The input sentence should be in written Sinhala format.
ii. The input sentences should terminate with correct punctuation mark. Such as question mark and full stop.

2.1.2 Overall architecture of tagger

Overall architecture of the proposed tagger, which is a two-step process that first runs through the tagged corpus and extract the linguistic knowledge. Then it runs through the row text inputs and generating the best tag sequence for the sequence of input words based on the knowledge that gathered from the corpus.

Lexical Parser: Checks boundary conditions of each sentences and words as defined in the lexical rules, and prepare for Tokenizing and Pre-processing. Tokenization run through the tagged corpus, separate out the words and tags,

Training: The next important step is training the tagger. The training method we describe here is based on supervised learning approach. It runs on the corpus, makes use of tagged data and estimates the probabilities of transition, P(tag | previous tag) and observation likelihood P(word | tag) using naïve bays classifier .

2.1.3 Use Case Scenario

Use Case Scenario Input Sinhala Sentence	
Use case ID	01
Use Case name	Input Sinhala Text
Actor	User , system
Pre-Condition	System is running, arithmetic question is received from the user and input Sinhala text in Sinhala font
Success Scenario	1. User enters Sinhala text to be analyzed in Sinhala font on the given interface. 2. User press on the Submit button on the interface.
Exceptions	1.a If user entered another language font system prompt an error message.
Post-Conditions	

Table 2 Use Case Scenario 1

Use Case Scenario POS tagging for Sinhala	
Use case ID	02
Use Case name	POS tagging for Sinhala
Actor	User , system
Pre-Condition	System is running, arithmetic question is received from the user in Sinhala language.
Success Scenario	1. The use case starts when the system received a arithmetic question 2. Token extraction from given question. 3. Extracting nouns, verbs, adjectives …. etc. 4. Allocate tag for each words.
Exceptions	1.a If user entered other language font system prompt an error message.
Post-Conditions	System has successfully identify word category

Table 3 Use Case Scenario 2

2.1.4 Non- functional requirements

Usability Simple user interfaces have used. Therefore anyone who wants to solve arithmetic problem in Sinhala language. They can use this system despite having a basic knowledge about handling computers.

Performance

If the user enters a valid input to the system, within a few seconds after clicking the analyze button the generated output will be displayed to the user.

Extendibility

The proposed system is developed in a way that it can be extended by others in the future, for example if more complex arithmetic problems such as division and multiplication.

Availability

A guidance to use this system would be freely available on the system itself. Therefore any user who wishes to try it out can use it.

Reliability

Reliability is the ability of the system to carry out its normal procedure with minimum failures. To provide reliability to the application testing will be carried to identify and fix each and every possible bug in the system before releasing the final product. Each and every individual component will be tested and finally the integrated system will also be tested under different conditions.

Maintainability

The system corpus needs to be updated with new words and rules in. In order to provide most efficient response to user requests we have to maintain the corpus. By updating corpus with most recent information we can do that. That is another major concern in our system because without maintaining the system team can’t achieve system goals.

Security

Unauthorized access to the corpus data is restricted.

2.2 Implementation and Testing

Implementation details about key word identification. This describe the details of the approaches and tools we used in the implementation as well as the models we used , various tools that we were used to manage the code , builds and code quality.

2.2.1 Technologies used

- Python (2.7. 1) and WinPython-64bit-3.4.3.7

Some reason to select python as the programming language

i. Python is very flexible language
ii. Huge stranded liberties
iii. Generally good quality documentation for standard library.
iv. Easy to learn and use
v. Python is free

Since in the project it is specially addressed that the python environment we also need to install NLTK .It is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WorldNet, along with a suite of text processing libraries for classification, tokenization, stemming, and tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.
Some of the major benefits of the WinPython is portable, runs on any windows with 2 GB Ram. And when we try to display Sinhala fonts on the python it would be a problem. But Win Python supported for Sinhala characters (both utf-8 and utf- 16)

For the training data set use Naïve bays classifier which has been shown to work with very large number of attributes (e.g. bag-of-words for text classification); and similarly for classes. Installed Naïve bays classifier using the pip command.

2.3 Testing

The performance of the tagger was measured, using three different versions of progressively upgraded tagging mechanisms. Test case 1 is the simplest form of the tagger that performs well only with known words and for simple sentence. Test 2 is a somewhat upgraded version that treats all known words tagged and each new word encountered in the tagging process. Test case 3 is the full version of the tagger that uses statistical technique to guess the best tag for unknown word by considering the context of surrounding words. Test case 6 tagger that performs well both known words and unknown words .

Test Cases

	Test case 1
Description	Tag one sentence without numbers
Sample input	මම පැන්සල් මිලදී ගත්තෙමි.
Expected output	මම _PRO පැන්සල්_NN මිලදී_VB ගත්තෙමි_VB
Actual output	මම _PRO පැන්සල්_NN මිලදී_VB ගත්තෙමි_VB

Table 4 Test Case 1

	Test case 2
Description	Tag one sentence with numbers
Sample input	මම පැන්සල් 2 ක් මිලදී ගත්තෙමි.
Expected output	මම _PRO පැන්සල්_NN 2_CD ක්_PP මිලදී_VB ගත්තෙමි_VB
Actual output	මම _PRO පැන්සල්_NN 2_CD ක්_PP මිලදී_VB ගත්තෙමි_VB

Table 5 Test Case 2

	Test case 3
Description	Tag two sentence with punctuation marks
Sample input	මම පැන්සල් 2 ක් මිලදී ගත්තෙමි. ඉන් 1 ක් මල්ලිට දුන්නේය .
Expected output	මම _PRO පැන්සල්_NN 2_CD ක්_PP මිලදී_VB ගත්තෙමි_VB . _SW ඉන්_PP 1_CD ක්_PP මල්ලිට_NN දුන්නේය_VB . _SW
Actual output	මම _PRO පැන්සල්_NN 2_CD ක්_PP මිලදී_VB ගත්තෙමි_VB

Table 6 Test Case 3

	Test case 4
Description	Tag three sentence with punctuation marks
Sample input	මම පැන්සල් 2 ක් මිලදී ගත්තෙමි. ඉන් 1 ක් මල්ලිට දුන්නේය . දැන් මා ලග ඉතිරි පැන්සල් ගණන කොපමණද ?
Expected output	මම _PRO පැන්සල්_NN 2_CD ක්_PP මිලදී_VB ගත්තෙමි_VB

	. _SW ඉන්_PP 1_CD ක්_PP මල්ලිට_NN දුන්නේය_VB . _SW දැන්_PP මා_PRO ලග_PP ඉතිරි_PP පැන්සල්_NN ගණන_NN කොපමණද_QW ?_SW
Actual output	මම _PRO පැන්සල්_NN 2_CD ක්_PP මිලදී_VB ගත්තෙමි_VB . _SW ඉන්_PP 1_CD ක්_PP මල්ලිට_NN දුන්නේය_VB . _SW දැන්_PP මා_PRO ලග_PP ඉතිරි_PP පැන්සල්_NN ගණන_NN කොපමණද_QW ?_SW

Table 7 Test Case 4

	Test case 5
Description	Tag three sentence with punctuation marks
Sample input	කමල් ලග පොත් 4 ක් තිබේ . තවත් පොත් 3 ක් අම්මා කමල්ට දුන්නාය. දැන් කමල් ලග ඇති මුළු පොත් ගණන කොපමණද ?
Expected output	කමල්_NN ලග_PP පොත්_NN 4_CD ක්_PP තිබේ_VB . _SW තවත්_PP පොත්_NN 3_CD ක්_PP අම්මා_NN කමල්ට_NN දුන්නාය_VB . _SW දැන්_PP කමල්_NN ලග_PP ඇති_PP මුළු_PP පොත්_NN ගණන_NN කොපමණද_QW ? _SW
Actual output	කමල්_NN ලග_PP පොත්_NN 4_CD ක්_PP තිබේ_VB . _SW තවත්_PP පොත්_NN 3_CD ක්_PP අම්මා_NN කමල්ට_NN දුන්නාය_VB . _SW දැන්_PP කමල්_NN ලග_PP ඇති_PP මුළු_PP

	පොත්_NN ගණන_NN කොපමණද_QW ? _SW?_SW

Table 8 Test Case 5

	Test case 6
Description	Test unknown words with Sinhala base word (මිනිසා)
Sample input	මිනිසෙක් මිනිසාට මිනිසෙකු මිනිසෙකුට
Expected output	මිනිසෙක් _NN මිනිසාට _NN මිනිසෙකු _NN මිනිසෙකුට_NN
Actual output	මිනිසෙක් _NN මිනිසාට _NN මිනිසෙකු _NN මිනිසෙකුට_NN

Table 9 Test Case 6

CHAPTER 03

3 Result and Discussion

3.1 Evidence

Even though features and methods have been presented, there should be substantial evidence that is able to prove the noted facts. Evidence plays the major role in software development industry while delivering the end product to user. Using provided evidences end user can get clear understand about the project, can identify the processes behind the development and implementation of the system and the quality of the end product. Evidence carrying the output of the final implementation according to the methodologies applies to gain those outcomes. To provide sufficient evidence there should be a proper verification and validation criteria. Therefore the system was put through unit testing, integrating testing, system testing and usability testing. While go through those testing areas developed system is capable of Solving most of the mathematical problem after identifying tagged words.

3.2 Discussion

A collection unannotated text was used for the purpose of testing the performance of the tagger. The model described by the estimated parameters has shown some interesting results when compared with to the size of the corpus.

The amount of data used in the training proses was not sufficient to identify the probability distribution of the large tag set defined. In order to avoid the possible consequences of the above situation an assumption was made when the performance is measured.

When the corpus size is around two thousand distinct words (2000) words, even though the unknown word percentage is 100%, the tagging error over all the words is reported below 40%.

To get the results, 100 sample sentences were used. The following list shows some sample sentences and the tagged words.

1. අමල් ලග පොත් 10 ක් තිබෙනවා . කමල් අමල්ගේ පොත් 4 ක් මිලදී ගත්තේය . අමල් ලග පොත් කියක් ඉතිරිද ?

   අමල්_NN ලග_PP පොත්_NN 10_CD ක්_PP තිබෙනවා_VB ._SW කමල්_NN අමල්ගේ_NN පොත්_NN 4_ CD ක්_PP මිලදී_VB ගත්තේය_VB . අමල්_NN ලග_PP පොත්_NN කියක්_QW ඉතිරිද_QW ? _SW

2. සමන් ලඟ රුපියල් 100 තිබේ . එයින් මල්ලිට රුපියල් 20 ක් දුන්නේය . සමන් ලඟ ඉතිරි මුදල කොපමණද ?

   සමන්_NN ලඟ_PP රුපියල්_ADJ 100_CD තිබේ_VB . _SW එයින්_PP මල්ලිට_NN රුපියල්_ADJ 20_CD ක්_PP දුන්නේය_VB . _SW සමන්_NN ලඟ_PP ඉතිරි_NN මුදල_NN කොපමණද_QW ?_SW

3. වට්ටියක සුදු මල් 12 ක් සහ රතු මල් 10 ක් ඇත . තවත් සුදු මල් 3 ක් එයට දැමූ විට මුළු මල් ගණන කොපමණද ?

   වට්ටියක_NN සුදු_ADJ මල්_NN 12_CD ක්_PP සහ_PP රතු_ADJ මල්_NN 10_CD ක්_PP ඇත_VB . _SW තවත්_PP සුදු_ADJ මල්_NN 3_CD ක්_PP එයට_PP දැමුවිට_VB මුළු_ADJ මල්_NN ගණන_NN කොපමණද_QW ? _SW

4. මගේ ලඟ දොඩම් ගෙඩි 50 ක් තිබේ. අක්කා තවත් 20 ක් මට දුන්විට මුළු දොඩම් ගණන කොපමණද ?

   මගේ _PRO ලඟ_PP දොඩම්ගෙඩි _NN 50_CD ක්_PP තිබේ_VB . _SW අක්කා_NN තවත්_PP 20_CD ක්_PP මට_PRO දුන්විට_VB මුළු_ADJ දොඩම්_NN ගණන_NN කොපමණද_QW ?_SW

5. ටැංකියක මාළු 10 ක් සිටි .එයට මාළු 4 ක් දැමුවිට මුළු ගණන කොපමණද ?

ටැංකියක_NN මාළු_NN 10_CD ක්_PP සිටි_VB . SW එයට_PP මාළු_NN 4_CD ක්_PP දැමුවිට_VB මුළු_ADJ මාළු_NN ගණන_NN කොපමණද_QW ? _SW

The accuracy result of the 100 sample sentences. The experimental result shows 74% of the sample is tagged perfectly and 12 % of the sample is basically OK. Therefore, the function

Gives nearly 86 % accuracy of the tagging.

Accuracy results

Test case	**Sentences**
Perfect tagging	74
Basically ok	12
Meaningless	11
Error	3

Table 10 Accuracy results

CHAPTER 04

4 CONCLUSION AND FUTURE WORK

4. 1 Conclusion

During the past decades, hundreds of arithmetic problem solving systems have been developed all over the world. Most of these systems have been developed for English language. All of these approaches and successful systems have been discussed in the first chapter. In addition to the above, available arithmetic problem solving systems were also discussed in the first chapter. It is very unfortunate that most of the people does not have enough knowledge about English language so they are not capable of using that kind of systems. We discussed research gap of the developed system in the first chapter.

Next chapter cover up methodology and implementation, testing and findings. To solve arithmetic problem in Sinhala language Accuracy is very important. The accuracy of tagging depends on the accuracy of the input text as well as on the software.

The final objective is to “Evaluate the system”. Arithmetic problem solving system has been evaluated through the three stages. Evaluation was conducted through the white box testing approach.

4.2 Future work

Further work is still to be done in several directions. Some of this corresponds to development of resources such as increase the size of corpus, while others refer to specific details of implementation and tuning.

This method may fail due to a small amount of training text. To address the above problem, the probability of the unknown word tags can be approximated by the less probable word tags i.e. tags of the word occurring only once or twice. The problem can also be studied by considering the valid linguistic suffix of the word instead of considering the first few characters of a word.

POS tagging algorithm. Further, features can be investigated in the Naïve bays classification framework to improve the tagging accuracy. However effect of inclusion of more specific features (i.e. is previous word belongs to a particular set, is next word is from a particular set) instead of the generic features can be studied in Naïve bays classification based framework of POS tagging.

All in all, the development of a machine learning based good accuracy POS tagger requires a large amount of training data. The future work also includes the development of a large amount of annotated data which can be further used for training the system. The present tagger can be used for the initial annotation and the errors can be manually checked which otherwise a very difficult task to annotate large amount of corpus. We also plan to explore some other machine learning algorithms (e.g. Support Vector Algorithm and Neural Networks) to understand their relative performance of POS Tagging task under the current experimental setup.

REFERENCES

[1] Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy Lushan Han1, Tim Finin1;2, Paul McNamee2, Anupam Joshi1 and Yelena Yesha1 1 Computer Science and Electrical Engineering ,University of Maryland, Baltimore County , 2 Human Language Technology Center of Excellence , Johns Hopkins University 29 December 2011

[2] "Learning to Solve Arithmetic Word Problems with Verb Categorization" (by Mohammad Javad Hosseini (University of Washington), Hannaneh Hajishirzi(University of Washington), Oren Etzioni(Allen Institute for AI), and Nate Kushman(Massachusetts Institute of Technology)

[3] Automatic Key phrase Extraction based on NLP and Statistical Methods Martin Dostal and Karel Je_zek Department of Computer Science and Engineering, Faculty of Applied Sciences University of West Bohemia, Plze_n, Czech Republic, fmadostal, jezek kag@kiv.zcu.cz

[4] Automatic Keyword Extraction From Any Text Document Using N-gram Rigid Collocation Bidyut Das, Subhajit Pal, Suman Kr. Mondal, Dipankar Dalui, Saikat Kumar Shome International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-2

[5] Sinhala Ontology Generator for English to Sinhala Machine Translation conference paper · Budditha Hettige General Sir John Kotelawala Defence Unive…George Rzevski The Open University (UK) January 2014

[6] A. R. Weerasinghe, Available: http://www.ucsc.cmb.ac.lk/People/rw/index.htm

[7] http://www.cplusplus.com/forum/general/ [Accessed : 22nd February 2016]

[8] http://texlexan.sourceforge.net/ [Accessed: 21st August 2013].

[9] A Stochastic Part of Speech Tagger for Sinhala Dulip Lakmal Herath A.R.Weerasinghe Language Technology Research Centre University of Colombo School of Computing Sri Lanka dherath@webmail.cmb.ac.lk arw@ucsc.cmb.ac.lk

[10] S. N. Kim, T. Baldwin, M. Kan, “Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction”, Proceedings of the 23rd International Conference on Computational Linguistics, Beijing Coling, 2010, pp. 572–580

[11] Learning to Solve ArithmeticWord Problems with Verb Categorization Mohammad Javad Hosseini1, Hannaneh Hajishirzi1, Oren Etzioni2, and Nate Kushman31fhosseini,hannanehg@washington.edu,2OrenE@allenai.org, 3nkushman@csail.mit.edu 1University of Washington, 2Allen Institute for AI, 3Massachusetts Institute of Technology

[12] A step towards a Natural Language Programming Tool (NLPT)
K. J. A. Perera1, K. A. I. Kuruppu2, M. P. Gamage3, J. A.P.B. Jayakody4, K. S. G. S. Gunasekara5, G. Nuwan Kodagoda6 Sri Lanka Institute of Information Technology, Malabe Campus, Sri Lanka

[13] A Math-Aware Search Engine for Math Question Answering System
Tam T. Nguyen, Kuiyu Chang, and Siu Cheung Hui Nanyang Technological University 50 Nanyang Avenue, Singapore 639798 nguy0080@e.ntu.edu.sg, {askychang, asschui}@ntu.edu.sg

[14] “Improved Automatic Keyword ExtractionGiven More Linguistic Knowledge” Anette Hulth Department of Computer and Systems Sciences Stockholm University Sweden hulth@dsv.su.se

[15] al, A. Bharathi et. Natural Language Processing : A Paninian Perspective. New Delhi : Prentice-Hall of India.

[16]http://docplayer.net/15520348-Unknown-words-analysis-in-pos-tagging-of-sinhala-language.html [Accessed : 29th February 2016]

[17] Daniel Jurafsky and James H. Martin, Speech and Language Processing, Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Person Education Inc (Singapore) Pte. Ltd., 5th Edition, 2005

[18] UCSC/LTRL Sinhala Corpus, Beta Version April 2005, downloaded from http://www.ucsc.cmb.ac.lk/ltrl/?page=downloads

[19] UCSC Tagset, downloaded from http://ucsc.cmb.ac.lk/wiki/index.php/file:ucsc_tagset.pdf.

[20] Tetsuji Nakagawa, Taku Kudoh and Yuji Matsumoto, "Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines". Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara, Japan

[21] http://mylanguages.org/learn_sinhala.php [Accessed : 20th February 2016]

GLOSSARY

NLP – Natural language Processing

POS – Part Of Speech

QA – Question answer

UCSC – University of Colombo

NLTK – Natural Language tool Kit

APPENDIX

Appendices