CONTEXT SENSITIVE TRANSLATION OF ENGLISH WORDS AND PHRASES TO SINHALA IN WEB CONTENT

W. M. G. M. Alles

2024

Context sensitive translation of English words and phrases to Sinhala in Web content

W. M. G. M. Alles
2024

Context sensitive translation of English words and phrases to Sinhala in Web content

A dissertation submitted for the Degree of Master of Computer Science

W. M. G. M. Alles
University of Colombo School of Computing
2024

Declaration

Name of the student:	**W. M. G. M. Alles**
Registration number:	**2019/MCS/004**
Name of the Degree Programme:	**Master of Computer Science**
Project/Thesis title:	**Context sensitive translation of English words and phrases to Sinhala in Web content**

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.

3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.

4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.

5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.

6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	**Date (DD/MM/YYYY)**
	23/09/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	**Supervisor 2**	**Supervisor 3**
Name	Dr. Ruvan Weerasinghe		
Signature			
Date	30/09/2024		

ACKNOWLEDGEMENTS

It was a challenging adventure to conclude this research project in a year, but it was also a tremendous learning experience. Without the generous assistance, direction, and inspiration of numerous people who stood by my side, I would not have been able to finish my research. I want to say thank you to each one of you.

I would first like to thank my supervisor, Dr Ruvan Weerasinghe, for his guidance, and support and for pushing me to do my best Master's Thesis, from the idea initiation to the final submission. Furthermore, I would like to thank all the lecturers at the University of Colombo, School of Computing for their endless knowledge and guidance throughout the master's degree period.

Next, I would like to express my sincere thanks to all evaluators, domain, and technical experts, of this research for providing their expertise and knowledge throughout the research. Further, I would like to extend my gratitude to my friends and company colleagues for their valuable feedback and support during the project.

Finally, I would like to express my loving gratitude to my family, who has been my pillars of success throughout my life.

ABSTRACT

Online resources are most popular knowledge acquiring technique available in today’s world. Several authors around the world publish articles which are in different languages in different literacy levels based on the author of the content. However, the readers may not get the full benefit of the online published resource such as article or documentation because an ordinary reader which makes a gap and uncomfortable when reading the particular web page content if it is include many difficult words which the reader does not know. In any language there is an ambiguity present but they are resolve based on the context of the sentence. Homonyms words are having different meanings according to the context for same word. If the context ambiguity is present for a sentence, the readers may frustrated or understand wrong meaning according to his literacy level. Because of this language barrier, readers may avoid reading some of the good contents as well.

There are several online translation tools and plugins are available around the world. From this thesis, design and implementation of browser plugin is presented for English based web pages /sites and it is focused on the Sri Lankan people. Non- English user community has been left out because the author assumes that readers having at least mid-level understanding of English language are reading the English based web sources to acquire knowledge. As mentioned earlier there are many online translations tools available, but these are capable of translating whole sentence to the reader according to the preferred language they select. These types of tools do not help readers to learn particular language. But by using the proposed system the readers can mouse-hover to the particular English word that he/she wants to get the Sinhala definition while learning the language unintentionally.

Natural language processing with Bidirectional Encoder Representations from Transformers (BERT) is used with machine learning approach in order to make this system success. In this thesis a computer enabling browser plugin with good network connection is only the simplest requirement the particular user need. Google Colab is used to implement the system prototype. User mouse-hovered word with the sentence will input to the system and based on the training model the system present it predicts the Sinhala definition and display as a tool tip. With this approach the reader not frustrated if the context ambiguity is present in the web page because within few milliseconds it will display the output and they can enjoy and acquire knowledge from the content and at the same time they can learn the English language unintentionally.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Chapter Overview

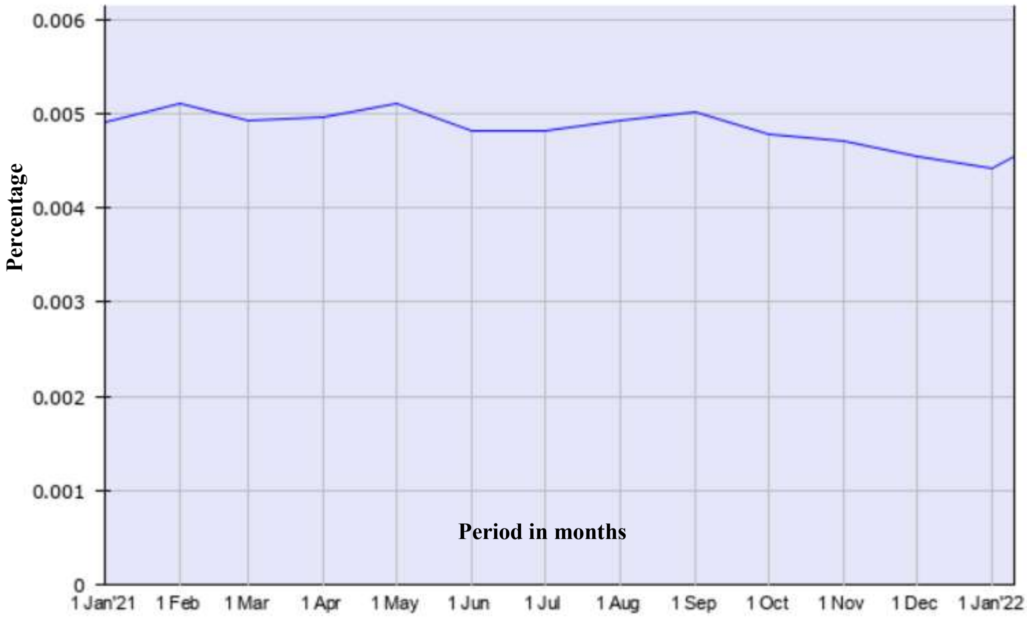
This chapter describes the background study related to the project. Starting from the introduction it discuss about the necessity of such translation tool when referring to the online resources. Under the literature review it discusses different approached and techniques that used to address similar problems with other important approaches, methods, tools and techniques.

1.2 Introduction

There are several ways people can acquire knowledge. Reading a web material or reading a physical book, watching videos and learning from a lecturer in a physical or online classroom are some of them. Each and every way specified above has different aspects. For example reading a web material is different than reading a physical book. There are many things that can perform when reading web material such as select a specific word, copy and paste it in an online dictionary available to get the meaning. But when reading a physical book if there is a difficult word, that person should type it in an online dictionary or search it using a physical dictionary. It is clearly understandable fact that, some ways have easy methods while some have time consuming and need a considerable amount of effort to put to accomplish the task. Unlike in the past, people now commonly rely on online resources for their daily activities. As of January 2022, statistics show that 63.6% of websites use English, making it the most dominant language online. The second most common language is Russian, used by 6.9% of websites. Compared to the English language, usage of Russian language is very small. It is a clear indication that most of the online resources are available in English language. According to the usage statics of Sinhala, Sinhalese (See Figure 1.1), it used by 0.1% in web content over the other language available in the world.

When consider about the Sinhala language, in Sri Lanka there are several webs sites use Sinhala with user preference. Hirufm.lk, Hirunews.lk, Gossiplankanews.lk, Adaderana.lk and Lankadeepa.lk are some of them. Most of the above sites have options to view the content in three languages Sinhala, English and Tamil which make the user to view the content which they familiar.

Figure 1.1: Usage statistics of Sinhala, Sinhalese as content language on the web



Anyone can agree that Sri Lankans not always refer web sites or web pages which are hosted by Sri Lankans because they may refer web sites or web pages hosted by any other country as well. As discussed earlier, most of the web pages or sites are available in English language. So they can acquire knowledge by referring those materials.

1.3 Problem domain/ Background

There are several different languages around the world. People may fluent in one or many languages. When carefully analyze a language, every language there is different relations in the senses of words. Relations can be mainly categorized into three, namely: meaning, written form and spoken form. For example there may be words with same meaning and written form or same written form and spoken form likewise there are different combinations. Below is an example of some words in English language categorized into above three types (See Table 1.1).

1.3.1 Homonyms

According to the above Figure there are several relations but this research mainly focuses about the homonyms. Homonyms are words that have the same spelling and pronunciation but have different meanings. The prefix homo- which means “the same” and the suffix –nym stands for “name”. Therefore a homonym is a word that has at least two different meanings, even though all uses look and sound exactly alike (See Table 1.2).

If a person is having less vocabulary, may not understand the correct meaning of the content. Therefore they face difficulties specially when referring to the several sources of particular language which is not there mother tongue. For this research, consider about the online web content as a source because a person is mostly refer them in their day to day life.

Table 1.1: Relations between senses

Name of Relation	*Spoken form*	*Written Form*	*Meaning*	*Examples*
Synonym	different	different	same	**happy, glad, joyful, blissful**
Homophones	same	different	different	**alms** (almsgiving) **arms** (hand)
				whole (all/ everything) **hole** (hollow / cavity) **hall** (lobby / building or large room used for meetings)
Homographs	different	same	different	**bass** (fish) **bass** (music)
Homonyms	same	same	different	**bat** (a nocturnal flying mammal) **bat** (use in sports Cricket/ baseball)
Orthographic Variants	same	different	same	**Color** **Colour**
Phonetic Variants	different	same	same	**either** /iy dh er / /ay dh er/

Table 1.2: Some of the Homonyms in English language

Homonym	*Meaning 1*	*Meaning 2*
Address	To speak to	Location
Bark	A tree's out layer	The sound a dog makes
Fair	Equitable	Beautiful
Mean	Average	Not nice
Rock	A genre of music	Stone

1.3.2 English Language

English is a West Germanic Language which originating from England. It is considered as a universal language. People from non-English speaking countries consider knowing English language is an additional advantage for their academic and professional carrier. Therefore they try their best to learn English language as it common way to communication. The Latin script which is modern English alphabet contains 26 letters which also have capitals from. There are 5 vowels and the pronunciation of vowels in English makes varieties of English. The structure of the English sentence is form by Subject, Verb and Object format. For academic standards there are two national standards are present namely British English and American English. They differ by spellings, grammar and vocabulary (See Table 1.3).

For example,

Table 1.3: Differences between British English and American English

Type of the difference	*American English*	*British English*
Spellings	defense, offense, license	defence, offence, licence
	color, behavior, mold	colour, behaviour, mould
Grammar	‘gotten’ as the past participle of ‘get’	‘got’
Vocabulary	mailbox	postbox
	apartment	flat

Translating from English language to German or French is quite easier than translate to Sinhala because French and German languages have close relationship among each other since all of them are based on a similar alphabet.

1.3.3 Sinhala Language

Sinhala is derived from ancient Brahmic and is an Indo-Aryan language primarily spoken by Sinhalese people of Sri Lanka. It is written using Sinhala Script and written from left to right. To represent different sounds diacritics called “පිලි” (“ඇලපිල්ල”, ”ඉස්පිල්ල”, ”පාපිල්ල”, ”කොම්බුව” etc…) are used in before, after, above or below the base-consonant. It does not have upper or lower case letters but it has අල්පප්‍රාණ (ක, ග, ච, ට, ත, ප etc...) and මහාප්‍රාණ (ඛ, ඝ, ඡ, ඨ, ථ, ඵ etc...) letters which have more weight when pronounce. Structure of Sinhala sentence is form by Subject, Object and Verb format.

The most significant difference in Sinhala letters over other languages are Sinhala Letters are curlicues. These Sinhala letters are ordered into mainly two sets namely Pure Sinhala and Mixed Sinhala. Pure Sinhala (ශුද්ධ සිංහල), which is a subset of Mixed Sinhala (මිශ්‍ර සිංහල) alphabet. The complete Sinhala script consists of about 60 letters, 18 for vowels and 42 for consonants. However, only 57 (16 vowels and 41 consonants) are required for writing Pure Sinhala. In Sinhala language the subject has to agree with the tense, gender and the singular and plural (form) of the verb. For example:

1. “He ran for the train” is translated to “ඔහු දුම්රියට දිව්වේය”.
2. “She ran for the train” is translated to “ඇය දුම්රියට දිව්වාය”.

3. “I ran for the train” is translated to “මම දුම්රියට දිව්වෙමි”.
4. “They ran for the train” is translated to “අපි දුම්රියට දිව්වෙමු”.

The verb “ran” in four sentences explained above translated to “දිව්වේය″ , ″දිව්වාය” , “දිව්වෙමි” and “දිව්වෙමු” respectively. First two sentences are translated “දිව්වේය″ and ″දිව්වාය” because of the gender specified in the sentence. The last two sentences are translated to “දිව්වෙමි” and “දිව්වෙමු” because of the singular and plural from of the sentence. But in English language, single word ”ran”(Verb) used for above four sentences. Unlike English any written Sinhalese word is pronounced the same way it is written but there are several variations, accents, words of Sinhala use in several parts of Sri Lanka.

According to the Sri Lanka’s Department of census and statistics in 2011 the population was made up of Sinhala (82%), Tamil (9.4%), Moor (7. %), Burgher (0.2%) and others (0.5%). The literacy rate of Sri Lanka based on national statistics placed at 91.1% and majority of the Sri Lankans speak the national languages of Sinhala and Tamil (81.8% Sinhala and 14.9% Tamil). Most of the Sri Lankans use English as their second language except Burgher and Malay. As most of the Sri Lankans’ mother tongue is Sinhala they tend to use Sinhala when reading, writing, listening and speaking.

As discussed above there are very few websites in Sri Lanka is capable to produce Sinhala content in their websites. But for educational or any other purpose people need to refer to the online resources which are not viewable using Sinhala language. Each any every online resources have different language literacy level because the author’s English language literacy is directly reflects to its content. This language barrier become is problematic not only Sri Lankans but also non-native English people over the world.

1.4 Problem Definition

During the Covid19 pandemic situation people tend to refer online materials for communication unlike early days. Most of the web materials are in English language. So as a Sri Lankan there are only few websites available in Sinhala language but most of them are in English. Therefore if that person is not good at English will have a problem when translating a sentence to Sinhala. There are some plugins are available to translate the whole sentence or a specific word in Sinhala but with the homonyms of English as discussed above, those tools may not suggest the correct word based on the context. Therefore, the proposed system enables users to hover the mouse over a specific word, and a tooltip will display the appropriate Sinhala translation for that word, considering the context of the sentence. The proposed system mainly focus on web sites or web pages in English language where most of the Sri Lankans acquire the knowledge by referring those materials therefore the non-English user community is not consider for this approach. Existing translation solutions make it easy to translate the whole sentence to reader’s native language but when carefully considering those tools it unknowingly blunt the readers’ ability of learning a particular language. Therefore it is good to have a solution which not only bridges the gap of translating word to correct meaning according to the context but also help readers’ gradually learn English language. The approach only translates the mouse-hovered English word which he/she do not understand to Sinhala language. The solution is considered about the hard owing to multiple senses of English words and that require a context-sensitive translation of words.

1.5 Research Motivation

As discussed above currently available tools for English to Sinhala translation will translate the whole sentence. Some are capable of giving all the meaning for that particular English word if it is a homonym. Sometimes people need to type or copy and paste them in an online available dictionary and get the meaning. Those are time consuming and need several steps to accomplish a task. Therefore plugin will help user to stay on the current web page and using mouse-hover to specific word they can see the exact meaning of that word based on the context even though that word is a homonym. There is a big advantage when user can get by using this plugin. That is, it will also help user to increase their vocabulary unknowingly. If the tool is translating a whole sentence user will read the whole translated sentence in Sinhala but he may not even read that sentence in English and identify the difficult word that he may don’t understand. But using this time to time he may mouse-hover to that word and get the correct meaning. This will help user to read the word and think whether he knows the meaning. If he doesn’t know then he will mouse-hover to that word. This will help user to learn and buildup their vocabulary unintentionally.

1.6 Research Gap

Most of the research has focused on translating English sentences or words to Sinhala using rule-based techniques that incorporate syntax and semantic structures. The EnSiTip plugin, which also provides translations via mouse-hover, presents multiple meanings for a word without considering the context of the sentence. Unlike English language it is difficult to find corpus for Sinhala word with their meaning for several different sentences. Because of that, most of the researches based on rules. Therefore the proposed system will help to fill the gap by identifying the correct meaning of the word using machine learning approach. Creating a training model for specific word needs to have huge collection of words. Implementing a corpus then help other researches to add new sentences to the collections and have a good accuracy levels for words in future researches.

1.7 Research Contribution

There are several contributions can be get by the proposed system.

1.7.1 Domain Contribution

Few resources are available in the context of English to Sinhala or Sinhala to English translations. Therefore building a simple corpus which containing several sentences to train a specific word will be very useful if any other researcher in future. This corpus can be expanded time to time with words with related sentences. This will contributes to have a public corpus for Sinhala which includes set of sentences to train specific English word with its Sinhala translation.

1.7.2 Computer Science Contribution

Creating and enhancing the training model which then helps to get high accuracy level with the evolution of training models.

1.8 Research challenge

There are several challenges when doing this research. Building a training model is challenging task. In order to train a specific word for specific meaning, it needs huge collection of sentences. Not only for one word but also it needs to have several words with several sentences. Implementing final solution as a plugin is a challenging task because it should identifies the mouse-hovered word in web content and provides the solution accurately and quickly.

1.9 Research Questions

1. What are the major challenges when building a language model to get context sensitive Sinhala translation for a selected English word/phrase in the web content?
2. What attributes can build an accurate model?
3. What are the limitations of such a language model and how to improve?
4. How can the vast knowledge encoded in the English web be understood by native Sinhala speakers while not depriving them of the opportunity to learn English?
5. What kind of language model would best facilitate this goal?

1.10 Research Aim

It is very helpful for a reader if the difficult words in the content can be easily translate to their native language. Translating the whole content to native language is not a good way if the person is willing to learn the language. Therefore user can look for a meaning based on the content without translating the entire content. The aim of this research is support for Sri Lankans to read the English based web materials by understanding the content without any hesitation. The user can read online resources comfortably with fewer distractions while reading to get the correct meaning of a difficult word in Sinhala in that content. User only needs to mouse-over the word, so the context sensitive Sinhala meaning will be display in few seconds.

1.11 Research Objectives

The final outcome of this project which will help Sri Lankans to get the correct meaning of the word in Sinhala based on the content of the web-material. The aim of this study will be achieved by focusing on the following objectives.

- Analyze repositories available in online to get the English words and the meaning with the usage (several sentences) of the particular word.
- Building the translation dictionary which will be the main resource for this application.
  1. Collect words and their definitions in Sinhala by popular dictionaries such as MalalaSekara dictionary and other glossaries.
  2. Those Sinhala word collected from dictionary entries then stored in standard encoding (Unicode)
  3. Entries will be validate, clean and filter accordingly.
- Building the language model
  1. Use appropriate machine learning algorithms which will be used to get the correct definition of the selected word and display in Sinhala.
- Implement the application, which then helps users to get the definition in Sinhala by mouse-hovering to the specific word.
- The number of similar definitions in Sinhala to an English word or phrase may vary. Therefore to improve simplicity and the readability of the proposed system it considers only one definition.
- The main intention of this application is to reduce the time spent on searching the meaning of the word and not to translate the meaning of the whole sentence.

- By giving an accurate meaning the user can then easily understand it. Therefore mostly focuses on nouns and does not consider the tense of the sentence which will complex the implementation.

1.12 Project Scope

This application will implement with following features and functions:

- Focus only for the English based web pages.
- Users can get the correct meaning of the selected English word by mouse hovering to it.
- Selected word will be display the definition in Sinhala according to the content.
- Help user to display the meaning by mouse-hovering, so that it will display the definition within few milliseconds.
- User can learn English language while enjoy the reading, and improve their vocabulary unintentionally.
- Cover considerable amount of words to build the training model.

1.13 Resource Requirement

All the resources needed for the project are as follows.

1.13.1 Software Requirements

Table 1.4: English to Hindi translation of a sentence

Software Tool	*Purpose*
MS Excel	Building the training model and test model in csv formats
STS	For application development
Git Hub	Repository for source code management
Google Drive	Storing documents and csv files
MS Word	Document creation
Google Colab	Implementation

1.13.2 Hardware Requirement

- Processor – intel Core i5
- Ram – 8GB
- Hard disk(storage) – 25GB

1.13.3 Other skills required

- Creating a training and test model with collection of sentences.
- Preprocessing using NLP.
- Knowledge of evaluating the model with scientific methods.

CHAPTER 2
LITERATURE REVIEW

2.1 Literature Review

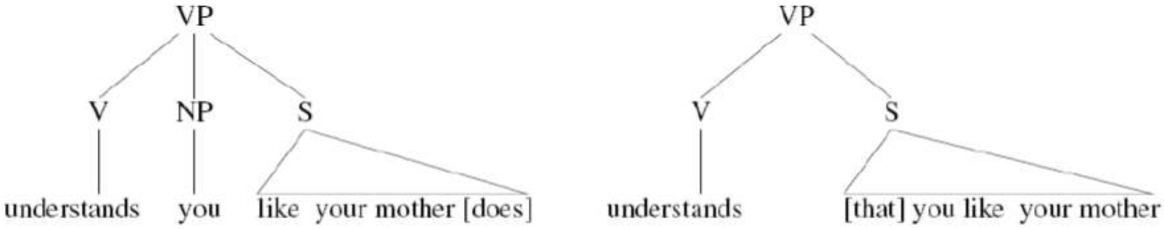
The focus of the literature review discusses the cross-language word sense disambiguation which is mainly useful for this proposed approach. There are two main areas to consider when implementing Sinhala translation tool for context sensitive words and phrases in the English language for online resources. First one is the translation dictionary and the other one is building a language model. Natural Language Processing and Machine Learning is two core concepts use for achieving areas mentioned above. Below it discusses the principles use for building the translation dictionary and the language model using several approaches, techniques and tools.

2.2 Tools and Techniques

2.2.1 Natural Language Processing (NLP)

This is recognized as a field of artificial intelligence. It will provide the accessibility to identify and understand the human language. Computers understand only the machine language. With evolution of technology, computers have the ability to understand text and spoken words in much the same way human beings can by processing large volumes of data as required. This has two directions. First one is how is the language is understand by the computer and the second one is how the machine is generate the natural language. For example, when the speech is input to the machine then it understand the phonetic, lexical, syntactic, semantic and lastly pragmatics then the machine should uses the reverse process when generating speech. NLP combines computational linguistics rule based modeling of human language with statistical, machine learning and deep learning models. Combination of these of these technologies enable computers to process human language in the form of text or voice data to understand its full meaning, complete with the speaker or writer’s intent and sentiment. Word sense disambiguation is a NLP task where the selections of the meaning of a word with multiple meanings through a process of semantic analysis that determine the word that makes the most sense in the given context. For example, word “mine” in “this pen is mine” refers to “belongs to me” while “gold mine” refers to “place where gold can be found”. There are several ambulation exists in many levels. If there is a disambiguation is exist in syntactic level, there are several parser trees can be generated (See Figure 2).

Figure 2.1: Syntactic level ambiguity



Therefore because of the ambiguous situations, identifying the correct context of a text is essential. The pragmatic analysis which is the last step of language generation process which is discussed above and it is a crucial part among the NLP components to overcome the drawbacks in the translation processes. There are several applications areas using the NLP approaches. Machine translation, text mining, Search engines like Google, text summarization, question answering systems for example IBM Watson are some of them. IBM Watson in Natural language understanding process it uses deep learning to extract meaning and Meta data from unstructured text data. Currently the chat bots are also capable of understanding and response to customers' issues which are raised through online applications and without human involvement chat bots can address some of the simple questions.

2.2.2 Machine Learning (ML)

Machine Learning is identified as a field of study which gives the capability for a computer to learn without being programmed explicitly. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classification or predictions. Training dataset including several sentences require to identify the specific word meaning. Language like python and their libraries are useful for providing the statistical results regarding the accuracy of them.

2.2.3 Machine Translation (MT)

This can be simply defined as a process when computer software translates text from one language to target language without human involvement. It can handle large amount of source and target languages that are compared and matched against each other by a machine translation engine. MT has ability to memorize key terms and reuses them wherever they might fit.

2.3 Cross- language word sense disambiguation

The lexical ambiguity remains one of the major problems for current machine translation systems. It states that in a sentence some words or phrases have different meanings based on the context of that particular sentence. One tool may translate that word to one translation which is having one meaning while the other tool translates the same word to another meaning. These types of ambiguity appear with this kind of lexical ambiguity of the word and the context. There are different methodologies have been investigated to solve the problem. Agirre and Edmonds (2006) and Navigli (2009) had done research regarding word sense disambiguation related algorithms and evaluation. Another research based on unsupervised word sense Disambiguation for twenty English nouns states that the sense label of which is composed of translations in different target languages (French, Italian, Spanish, Dutch and German). The basis of the Europarl parallel corpus is used to build the sense inventory and all translations of a polysemous word were manually grouped into clusters, which constitute different senses of that given word. Native speakers assigned a translation cluster(s) to each test sentence and based on their top three translations from the predefined list of Europarl translations considered in order to assign weights to the set of gold standard translations. These traditional monolingual word sense disambiguation converts into a cross-lingual word sense disambiguation and clearing the data acquisition bottleneck for word sense disambiguation in multilingual unlabeled parallel corpora involved with fairly little linguistic knowledge. In order to obtain the multilingual sense inventory, there are two approaches were used. One is to find all possible translations for our set of ambiguous focus nouns by performing word alignment on the parallel corpus and the other one is manually lemmatized all translations and meanings for clustered the resulting translations. Means of the cosine similarity between the training and test vectors used to inferred and compared to the matching training sentences with test sentence. (van Gompel and van den Bosch, 2013), solved the cross language word sense disambiguation by k-NN classifier.

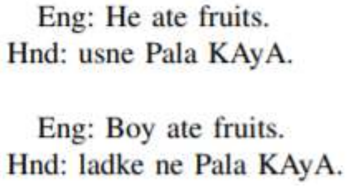
2.4 Existing approaches for machine translation

Human-assisted, Rule-based, Statistical, Example-based, Knowledge-based, Hybrid and Agent-based are some of the categories to classify Machine translation systems. Most of the categories use copra for the machine translation. Therefore these approaches are names as corpus based approaches. Statistical approaches are automatic learning methods and they need less time to build MT system but they need large parallel corpus for translation and the result not guaranteed to produce 100% correct translation. Therefore Rule-based approach is better than statistical approach when considering to the output result. When compared to the computer with human being, computers cannot store all the knowledge a person gain through his/her entire life. That is a major bottleneck for this type of methodologies.

2.4.1 One Indian language to another Indian language. (Chaudhury et al., 2010)

This approach uses Human-assisted methodology. This research (Anusaaraka) has been developed to translate Punjabi, Bengali, Telugu, Kannada and Marathi languages into Hindi. There is an English-Hindi Anusaaraka where it translates English text into Hindi. The approach and lexicon is general, but the system has mainly been applied for children's stories. This system uses Paninian Grammar model for its language analysis. This system focuses on producing more correct translation rather than giving a meaningful translation and reducing the language barrier by facilitating access from one language to another. Architecture of this system contains mainly Apertium which is for initial analysis the text and Anusaaraka for further processing. The process starts with de-formatting which get only the text by removing Html tags. Analyzer tokenizes the plain text and tagging process uses first-order hidden Markov model to choose corresponding lexical forms. Then lexical transformer gets source language lexical form and sends the matching target language lexical form. These are input to chunking process which then identifies patterns of lexical forms. For each lexical form deliver its target language surface from my morphological generator and then the re-formatter save output. After completing the initial process it sends to the CLIPS expert shell of the Anusaaraka for further processing.

Figure 2.2: English to Hindi translation of a sentence



Eng: He ate fruits.
Hnd: usne Pala KAyA.

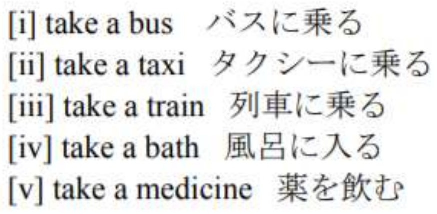
Eng: Boy ate fruits.
Hnd: ladke ne Pala KAyA.

Above is a translation from English to Hindi sentence by Anusaaraka (See Figure 3). The system architecture can be useful for two purposes. If the source language and target language are grammatically distinct, then Anusaaraka (English-Hindi) can be used. If the source language and the target language are grammatically similar then with minimum changes in target language part it can be useful for English-Telugu, English-Tamil pairs.

2.4.2 English to Japanese language. (Amano, 1987)

This system is translating open-domain written text by using morphological analysis, syntax analysis, translation word selection and structural transformation and morphological generation steps. Rules based approach with common word dictionary, a technical-term dictionary and user-defined dictionary is use by this system. Common word dictionary contains both English-Japanese and Japanese-English translation. Technical term dictionary contains 28 domains including computer, machinery and medicine . Japanese language doesn’t have spaces between words therefore morphological ambiguity is present when the input text is segmented into words and phrases. Second step is the syntactic analysis where augmented translation grammar which is context-free grammar used for parsing the sentences. Semantic analyzer is use to get the accurate meaning. There are different translation words available in any languages. In Japanese transitive verb “かけ る”(kake-ru) has different meanings. This type of words should be select carefully based on the context. Lexical transfer in tree conversion is useful for this. In the final step of generation phase, Japanese word order structure is considered to map the English words. This system produces the following output.

Figure 2.3: English to Japanese translation



[i] take a bus バスに乗る
[ii] take a taxi タクシーに乗る
[iii] take a train 列車に乗る
[iv] take a bath 風呂に入る
[v] take a medicine 薬を飲む

2.4.3 Hindi to English language. (Lavie et al., 2003a)

This system (XFER) is an approach translates from Hindi to English language. System used IIIT Morpher which is the morphology module for inflect words in Hindi. It considers gender, number and tense and uses Roman-WX for Romanized character encoding for Hindi. It contains 70 transfer rules, 58 verb sequence rules, 10 recursive noun phrase rules and 2 prepositional phrase rules. Except of the above manually written rules it consists of 327 rules which are automatically learned transfer rules. However this system is suitable for trained an extremely limited data scenario.

2.4.4 Pali to Sinhala Language. (Shalini and Hettige, 2017)

This system is used to translate Pali words to Sinhala language which was successfully translating simple Pali sentences to Sinhala. This can be used as a learning tool. Dictionary based approached is used in this system. It is mainly combination of three core components namely; Dictionary based translator, Pali morphological analyzer and the Sinhala morphological generator. Pali morphological analyzer is getting an input as Pali word and it shows the grammatical information and the root word. It is the main component among three component listed above. An affix spiriting approach is supports to identify the relevant root word in Pali. Pali dictionary consists of irregular words and it developed the root word indentation table. With the support of Pali Sinhala dictionary, Pali to Sinhala translator identifies based word for existing Pali word. Semantic issues and word level ambiguity are neglected and it shows an error message if the word is not available in the dictionary. The final phase of the model is the Sinhala morphology generator. Sinhala based word is generate appropriate Sinhala word by Sinhala Morphological generator. Limited number of rules used to generate Sinhala words. This system is not applied syntax level generation because it said that Pali and Sinhala languages are closely related to each other. This system is beneficial for the Buddhist monks in Pirivenas and any student who wish to learn Pali language specially in Dhamma schools

2.4.5 English to Arabic language. (Soudi, et al., 2007.)

This system used a mapping system for Arabic to intermediate representation. This mapping system contains three steps namely; selecting lexical items for each Interlingua concepts, mapping the semantic roles and mapping the semantic features for each Interlingua concept to appropriate syntactic feature in the feature structure. It has been test on 29 different structures and has produce good results.

2.4.6 Hindi / Bengali to English language. (Mandal et al., 2007)

Dictionary-based machine translation approach is the main methodology to develop this system where it uses cross language retrieval in the system. The queries in Bengali and Hindi translate to the equivalent English query out of Indian language topics. Phonetic translation system is used to overcome the limited coverage dictionary. 26,000 Hindi words and 9,000 Bengali words are used to build the dictionary. According to that it is not considered parts of speech information because of the dictionary limitation, improper stemming and the term is foreign word or a named entity. It stated that the Hindi word “rokanA” (to stop) has 20 translations which makes the average English translation per Hindi word in the lexicon were 1.29. That is because 14.89% of Hindi words have several translations.

2.4.7 Babel Fish (Yates and Sarah 2006)

This is a web based application initially called by AltaVista developed to translate single text, phrase or web pages (by proving the url) from one language to another. It can translate 13 languages and English, Chinese, German, Greek, Italian, Japanese, Korean, Portuguese, Russian and Spanish are some of them. This system use statistical machine translation approach. This approach works by analyzing parallel corpora that have already been translated from one language to another. For example, if it search “un perroquet rouge” in French every time “a red parrot” occurs in English. Then it stores these two phrases together in a “phrase table”. It also analyses large amounts of text in individual languages and memorize the frequency that certain words or phrases follow others. Then it build the language model identify different meanings of words when the input word have several meanings.

2.4.8 Example Based Translation (Weerasinghe, Premachandra 2008)

This system is primarily designed for use in the government sector to facilitate English-Sinhala translations using an example-based translation method. It employs a bilingual corpus for its knowledge base, enabling it to retrieve English sentences along with their corresponding Sinhala translations. The system identifies input phrases through intra-language matching. In terms of performance, it has achieved BLEU scores between 0.17 and 0.26 for 3-gram analysis when retrieving Sinhala sentences to determine the most frequently occurring Sinhala phrase in the dataset.

2.4.9 A Multi-Agent Solution (Hettige, Karunananda, and Rzevski 2016)

The system is developed as a multi-agent solution consisting of a six-agent swarm. It effectively addresses the morphological, syntactic, and semantic aspects of both the source and target languages. Each agent utilizes two language dictionaries and corpus ontology. Since people understand sentences by progressively reading words while considering syntax and semantics, this system follows the same principle. Built using the MaSMT framework, the system delivers more accurate translations for sentences of average length.

2.4.10 Translating Idioms- English-Sinhala (Bandara and R.R.T.K. 2019)

This research on idiom translation involved a sample of 10 Translation Studies students from the University of Kelaniya. The students were tasked with translating 10 common English idioms into Sinhala, and their methods of understanding and translating these idioms varied. The students primarily translated the idioms based on the literal meanings of the words. While idioms shared between both languages can be understood this way, some idioms have a figurative meaning that requires deeper knowledge. Therefore, translators need to have a thorough understanding of these idioms.

2.5 Approach for the project

In the modern era machine translation has received much attention because the decrement in time and human effort required for converting and translating words, sentences and paragraphs. Thus as discussed above there are several ways for machine translations. Each and every system described above uses different types of mechanism with different size of word counts to train the model. Most of them have been focused on Indo-European, Indo-Aryan or Sino-Tibetan families (Sjöberg). To get accurate result it should contains larger words collections. But that is not practical. For this project there are two tasks to be accomplished. Creating a translation dictionary and building a language model are two of them.

2.5.1 Translation Dictionary

This is the building block for the proposed system. This system is going to implement as an extension to the "EnSiTip" and it used Sinhala to English translation dictionary. It is constructed by gathering words and their definitions from the Malalasekara and other popular glossaries represented in standard encoding (Unicode). Those are formatted, validated, cleaned and filtered accordingly by group of experts. The user friendliness is enhanced by reducing the number of Sinhala definitions corresponding to each English headword. In order to achieve that the Occam's Razor principle was used.

2.5.1.1Occam's Razor principle

This principle states that "Entities should not be multiplied unnecessarily". There may be several theories which are competing for the same prediction, therefore, take the simplest one which makes everything better. William of Occam is the person who developed this concept and razor denotes that cutting off or shaving away the other possibilities.

Example 1, Assume in a rainy day corner of the room spilled out some water.

1st possibility with high probability: Should be leaky roof.

2nd possibility with less probability: Kid who spilled water in the room.

According to the above principle it omits fewer probabilities which help to make it simple rather than complex. That means water spilled because of the rain and the leaky roof.

Example 2, Assume car tire is flat when getting ready to leave.

1st possibility with probability: A nail stuck in the tire wall let the air out

2nd possibility with less probability: Someone slashed the tire.

According to the principle, it is more likely the tire gets slashed by nail.

By following the Occam's razor principle omit unnecessary meanings which make this very complex. Therefore according to that EnSiTip's translation dictionary is covered nearly

50,000 English words to produce better result for user. This translation dictionary was converted in a way which can be useful to the proposed system. There are many English words has different meanings based on the context of the sentence. Then the below translation dictionary was developed. There are more than 36000 English words are listed there.

2.5.2 Language model

2.5.2.1 Probabilistic model of language

It is attempt to characterizes, capture and exploit regularities in natural language. Determine the probability of a word in a sentence by analyzing the bodies of text data by using statistical and probabilistic approaches. This is more specifically called as statistical language models. Large amount of words are used to automatically determine the model's parameter in this statistical language models. Speech recognition, machine translation, context sensitive spelling correctors and next word prediction are the usage of this. It assigns probability to every word sequence which may be grammatical or not.

P [W1 W2 W3 … Wn].

Related: P (W5 | W1, W, W3) implies that conditional probability of W5 after W1, W, W3.

Bayes Rule/Chain Rule

P(X1,X2,X3,…,Xn) = P(X1)P(X2|X1)P(X3|X1,X2)…P(XN|X1,…,Xn-1)

For example, Sentence: I like red apple.

P(I like red apple) = P(I) P(like | I) P(red | I like) P(apple | I like red).

To estimate probability for text corpus,

P(I) = count ("I") / Total # of words

P(like | I) = count ("I like") / Count ("I") and so on.

2.5.3 N-gram language model

In the basic concept of the Natural language processing N-gram is using for the applications to build language models. For example this N can be 1, 2, 3, etc.. and according to that it defines as Uni-gram, Bi-gram, Tri-gram etc. This language models use some number of preceding words to makes predictions.

Simplest approximation: unigram

P(“I like red apple”) = P(“I”) P(“like”) P(“red”) P(“apple”).

Bigram

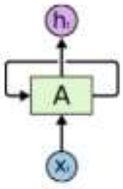
P(“I like red apple”) = P(“I”) P(“like” | ”I”) P(“red” | ”like”) P(“apple” | ”red”).

The size of the N should be considered based on the situation. Having N > 3 will give more accurate results. But the complexity is exponentially growing with the increase of N value. Data sparsity problem is common issue when handling the local dictionaries / vocabularies because there will be more missing values and not able to provide accurate results. Therefore Recurrent Neural Networks use to address the above issue.

2.5.4 Recurrent Neural Network (RNN)

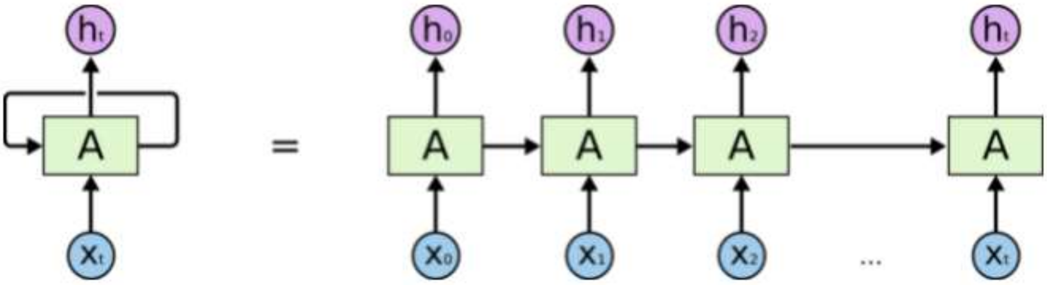
RNN can handle inputs having variable lengths. It is suitable for modeling the sequential data such as sentences in natural language. RNN contains loops, therefore same network copies again and again until it reaches to the successor. Figure 2.4 shows the RNN with its loops. “A” is a chunk of neural network which take the input “Xt” and produce output “ht” and the loop continuously pass the information from one step of the network to the next. With this loops it is same as having several copies of the same network.

Figure 2.4: Recurrent Neural Network



As discussed above RNN network can be unfold. Figure 6 illustrate how the above RNN network is unfold to multiple copies of the same network.

Figure 2.5: Unfolded Recurrent Neural Network



Because of these networks have chain like nature they have similarities to sequences and lists which are available in data structures. Currently, there are several researches and systems are built on top of this technique. Speech recognition, language modeling, translation and image capturing are some of the areas that RNN can apply. As above explained the RNN depends on the previous state, but with the long sequences grow over period of time it cannot handle larger among of data because the information which learnt by RNN will decay. This issue is called the “vanishing gradients problem”. Because of that the final result is not very accurate. Therefore several extensions were introduced to address this issue.

2.5.5 Long Short Term Memory (LSTM)

This technique is type of a RNN which address the issue of “vanishing gradient problem” in the RNN. LSTM can remember information for longer time period by keep track of previous events. It follows the same structure of RNN with introducing four new layers. Figure 2.6 and Figure 8 show the internal structure differences of RNN and LSTM network.

Figure 2.6: Internal structure of RNN

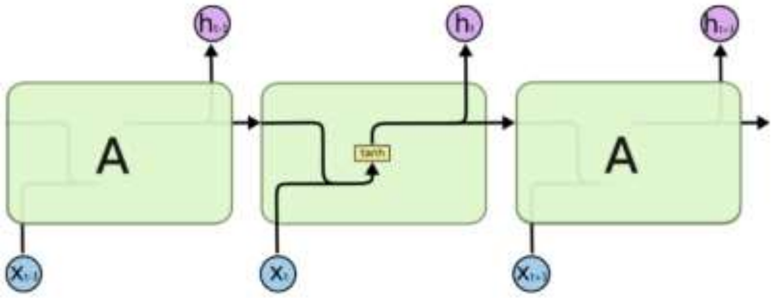
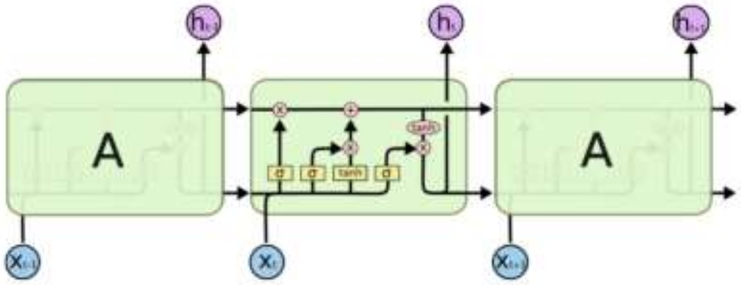
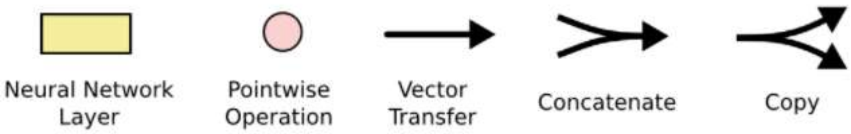


Figure 2.7: Internal structure of LSTM network



The symbols of the above two figures are denoted in the Figure 9.

Figure 2.8: Symbols used for internal networks



According to the Figure 2.7 the RNN structure has only one layer in its repeating module. But Figure 2.8 illustrates that the LSTM contains four layers in its repeating layer.

The internal structure and the behavior of the LSTM network are explained below with an example.

For a language model, if it wants to predict the next word for a sentence;

I grew up in China and I spent most of my life there but currently live in London. Therefore I speak fluent in <….> (Chinese)."

The first step of the LSTM network is to identify which information is going away from the cell state. "forget gate layer" is a sigmoid layer which make this decision. The output will be either 1 or 0 where 0 denotes by "completely get rid of this" and 1 denotes by "completely keep this". According to the example the cell state may include the gender of the present subject. Next step decide which new information needs to store in the cell state. This contains two parts; "input gate layer" which decide which part to be modified and second part is creating a vector for the new candidate values. According to the example in this step it adds the gender of the new subject to call state to replace the old one. After it updates the old cell state with the new cell state it actually drops the information about the old subject's gender. In the final step, decide what to be taken as output. According to the example it might output if the subject is singular or plural and decide the form of a verb should be produce.

As described above the process is continues and it remember values over arbitrary time intervals. Therefore LSTM is more accurate than RNN results.

2.5.6 Bidirectional Encoder Representation from Transformers (BERT)

This is used for keyword, semantic search to retrieve information and to produce vector based inputs from the words which will be uses in the Natural Language Processing models. It uses transformer which learn contextual relations between words in a text. Internally this transformer contains two separate mechanisms namely; an encoder to read the input text and a decoder produce a prediction for the task. For the language model only the encoder is necessary. For example, if the BERT is using then for the sentence "The man was accused of robbing a bank. The man went fishing by the bank of the river", the word embedding for "bank" would be different for each sentence.

BERT uses mainly two strategies pre-training and fine tuning. Pre-training strategy contains two tasks. They are Masked Language Model (MLM) and Next Sequence Prediction (NSP). As name implies in the MLM it is to mask some percentage of the input token randomly and predict them. Example, "I am eating an apple".

- 80% of the time: Replace the word with the [MASK] token e.g., I am eating an [MASK].
- 10% of the time: Replace with random word e.g., I am eating an orange.
- 10% of the time: Keep the word unchanged, e.g., I am eating an apple.

The other task mentioned in the pre-trained strategy is Next Sequence Prediction (NSP). In this method more than one sentence can be combined in a systematic way where it start the sentence with "CLS" and separate the sentence using "SEP". This approach is used to identify the relationship between the sentences.

For example,

- Sentence 1 -> "Kevin went to the supermarket".
- Sentence 2 -> "He bought bread and milk".
- Sentence 3 -> "Environmental pollution is increasing day by day".

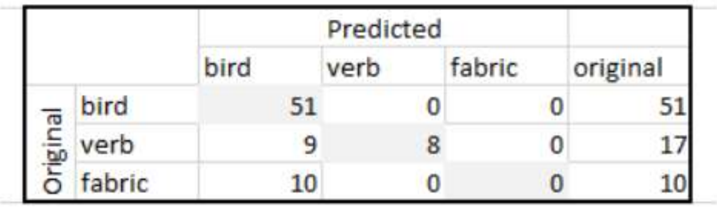
Using this method it will identify the sentence 1 and 2 are closely relate with each other while sentence 1 and 3 or 2 and 3 are not having any relationship. Second strategy in BERT is fine tuning. Fine-tuning is straightforward since the self-attention mechanism in the Transformer allows BERT to model many downstream tasks whether they involve single text or text pairs by swapping out the appropriate inputs and outputs. As discussed above there are many techniques and tools available and they are used in different approaches to build several systems. In early days B-gram language models used with immerge of the technologies latest trends are use. For this project BERT can be applied by training words using collection of sentences (Devlin et al. 2019).

There is an application built to identify the right meaning of the word "duck" where it contains four separate meanings based on the context and for the simplicity it categorized the verb "duck" and corresponding nous as the same meaning.

1. Bird / flesh of the bird 'duck'.
2. Lower head or body suddenly.
3. Durable closely woven usually cotton fabric.

The dataset contains 77 sentences contains the word "duck" and there are 50 sentences identified as animal (type 0), 17 identified as verb (type 1) and 10 identified as fabric (type 2). With the manual identification of the word "duck" and the actual was compared. The following figure shows the result gained from using BERT and the translation from the English-Hungarian translator.

Figure 2.9: Expected vs. Actual for the "duck" word



		Predicted			
		bird	verb	fabric	original
Original	bird	51	0	0	51
	verb	9	8	0	17
	fabric	10	0	0	10

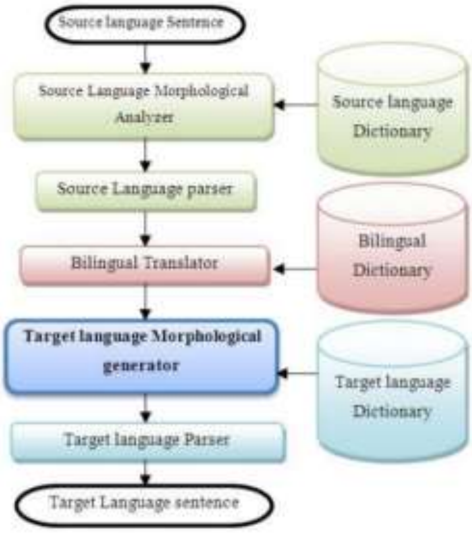
According to that the bird type is correctly identified and the 8 sentences are correctly identified as verbs. All the fabric type is incorrectly identified as bird. It produces 75.641% accuracy (Németh 2019).

2.6 Related work

2.6.1 Approaches for English to Sinhala translation

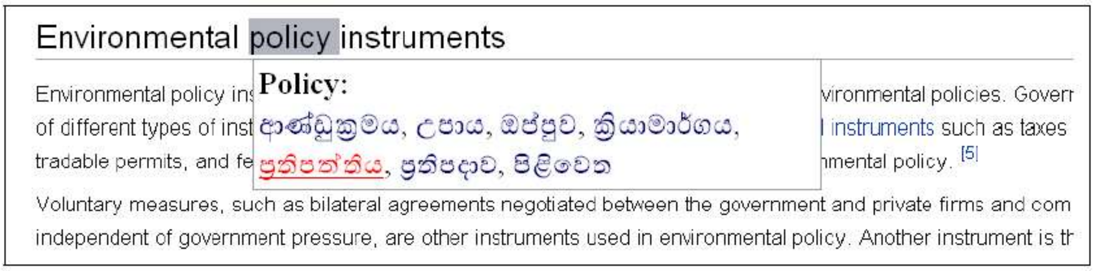
In Sri Lanka there are several universities contributed to develop machine translation system for Sinhala and Tamil languages. Sinhala Corpus, Parts of Speech Tagger, Optical Character Recognition system for Sinhala language and Sinhala text to speech system are some of them. The corpus based approach for Sinhala to Tamil machine translation system provided reasonable results for the evaluation by BLUE score. Moreover, many prototypes are developed for English to Sinhala machine translation. English to Sinhala translation system for weather forecasting have developed. It can translate simple sentences and works on the limited set of words with limited sentence patterns. This is a rule based and it has used paragraphs and sentence tokenization, simple parses, translators and Sinhala sentence generators for English to Sinhala translation. Bilingual Expert for English to Sinhala is another system which used rule based approach for implementing the system. It handles the language primitives such as person, gender, tense, number, preposition and subjectivity or objectivity. Moreover, it allows deriving all associated words from a given base word and thus it reduces the size of the Sinhala dictionary. Apart from that four lexical dictionaries are used. They are Sinhala dictionary, English dictionary, English-Sinhala Bilingual dictionary and concept dictionary (Hettige and Karunananda 2010). It contains seven modules which illustrates in Figure 2.10.

Figure 2.10: Generalize version for English to Sinhala translation



EnSiTip, is a word based English-Sinhala translation tool to support users who have less literacy level of English language to understand the web content (Wasala and Weerasinghe, 2008). It is a Firefox add-on and a user-friendly tool. Users can mouse-hover to a word which then appears a popup displaying the possible suggestion in Sinhala language for that particular word. Building the English-Sinhala dictionary was the core component for this system. It automatically de-inflects verbs and adjectives and from the tool user also capable to listen to the pronunciation of the English word. This system covers nearly 50,000 English words. According to the statics provided, there were 411 active daily users (Wasala and Weerasinghe, 2008.). Figure 2.11 shows how the EnSiTip will display the Sinhala meaning when mouse-hover to the word.

Figure 2.11: EnSiTip result from a web page


Environmental policy instruments

Environmental policy ins vironmental policies. Goverr
of different types of inst l instruments such as taxes
tradable permits, and fe nmental policy. [5]

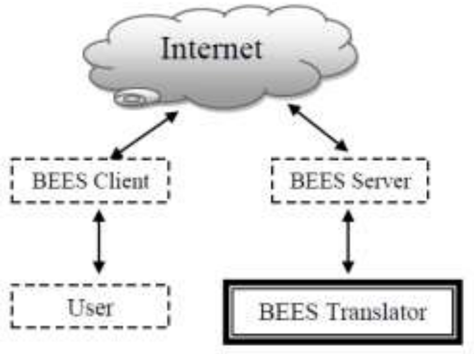
Policy:
ආණ්ඩුක්‍රමය, උපාය, ඔප්පුව, ක්‍රියාමාර්ගය,
ප්‍රතිපත්තිය, ප්‍රතිපදාව, පිළිවෙත

Voluntary measures, such as bilateral agreements negotiated between the government and private firms and com
independent of government pressure, are other instruments used in environmental policy. Another instrument is th

There is another system is developed to get the Sinhala translation for selected text. This system is built as an updated version of the English to Sinhala machines translation system; BEES (Hettige and Karunananda, 2010). This new system expanded the designed into three modules as BEES client, BEES server and BEES translator (Hettige et al., 2009). When the user selects the word BEES client read the highlighted text and send it to the BEES server.

This BEES server contains BEES translator which provides the meaning for the selected word. Three lexical dictionaries used for this system. They are English, Sinhala, and English-Sinhala bilingual dictionary. Experimental result shows that this system works with more than 80% accuracy and the human support is needed to test the system. The BEES client is capable to read selected text from document in any format such as word, pdf and html. Incomplete or incorrect text selection is an issue reported in the system because the rule-based top-down parser is less efficient to handle incomplete sentences than the grammatically correct complete sentences. Therefore system needed to improve to handle those issues.

Figure 2.12: BEES high-level architecture



Some of the works related to this project are discussed in the above. Currently, there is no exact system developed to get context sensitive Sinhala translation for selected word in English language. The proposed system is going to implement as an extension to the EnSiTip. Using the proposed system users can simply mouse-hover to a specific word and view the context sensitive meaning easily.

CHAPTER 3
METHODOLOGY

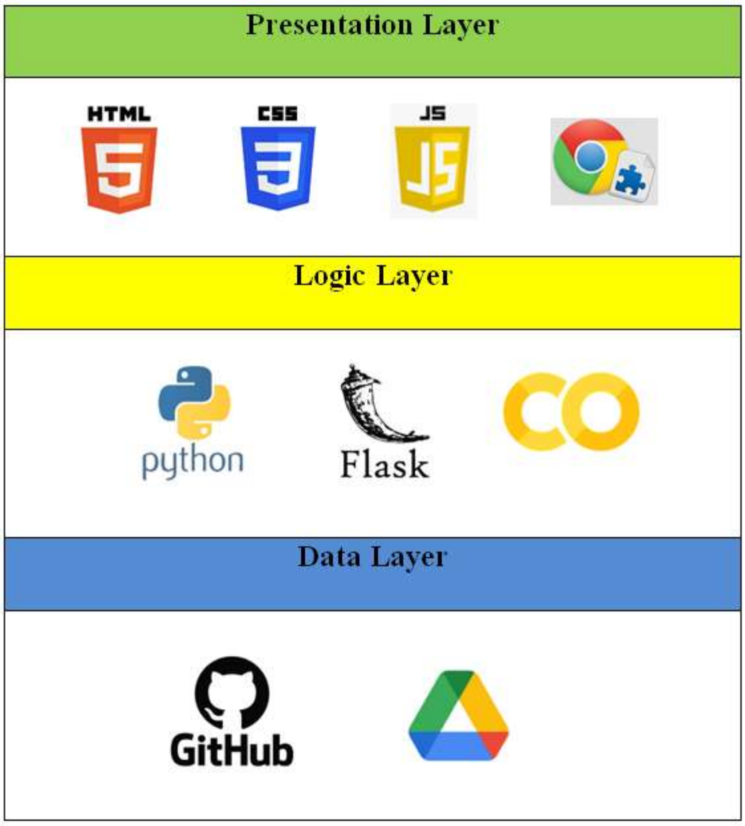
3.1 Introduction

The methodology is explained how the research problem is address and handle with the knowledge gain by the literature review. In the literature review mentioned there are several implementations for translating the whole English sentences to Sinhala language and EnSiTip is showing how the selected English word is translated to Sinhala word. As mentioned earlier the selected English word may contain different meaning based on the context of the sentence. So according to the EnSiTip implementation it showed the several Sinhala words (definitions) for selected English word. Because of this problem this research is based on newer approach to achieve the goal which means for selected English word it will show a relevant Sinhala word based on the content of the sentence. This chapter provides a comprehensive overview of implementation steps which have been carried out during the project to make it successful.

3.2 Technology Stack

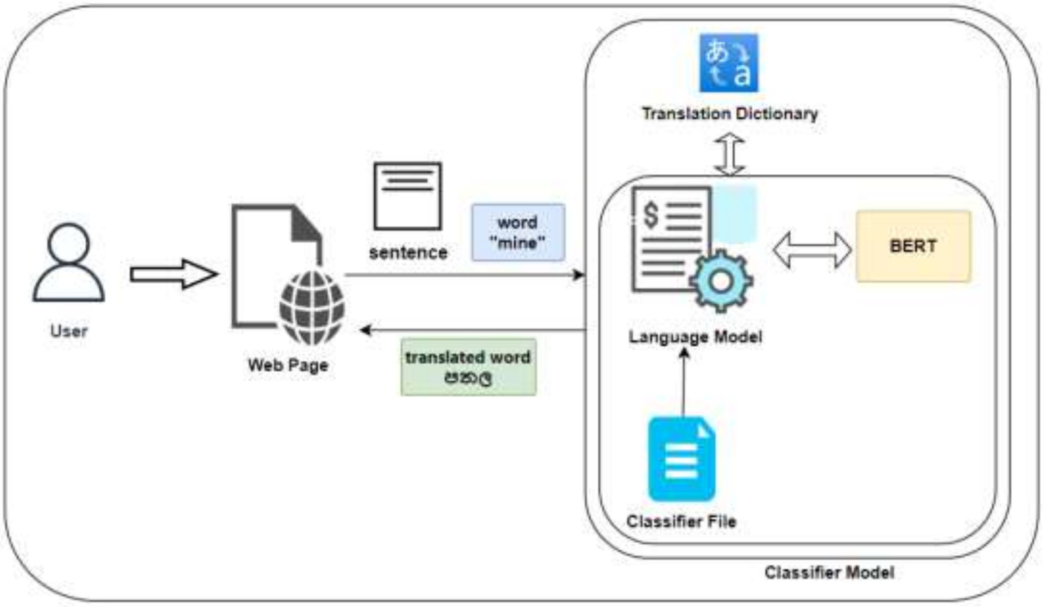
The selected technology stack for implementing different layers of the proposed system is outlined below.

Figure 3.1: Technology Stack



3.3 System Architecture

Figure 3.2: System Architecture



3.4 Representation of the Problem

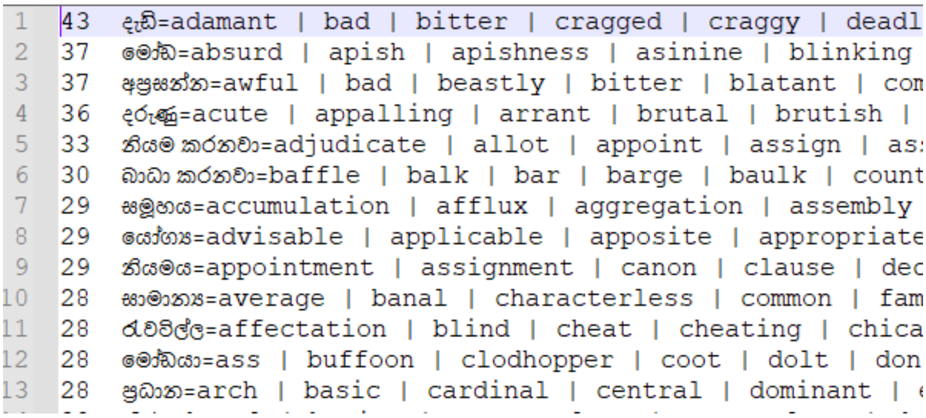
The aim of the study is to find a solution to display correct Sinhala word (definition) of a selected English word in web content. Collections of sentences are trained for specific English word which may have several classed according to the number of meaning of that specific word. This will be using machine learning algorithms to classify the sentences for correct classes. To classify the sentences according to the specific English word it uses BERT embedding. As for the initial step it needs to build a translation dictionary. The solution mainly contains two phases. First phase is to train a classifier to classify Sentences to the specific English word. And the second phase is to display the correct Sinhala word (definition) of the selected English word in the sentence.

The application has three separate sections. The initial step of the system is the user mouse-hover to a word in web content. Then the mouse-hovered word and the respective sentence feed to the classification model. Inside the classification model it contains language model. Language model is attempt to characterizes, capture and exploit regularities in natural language and determine the probability of a word in a sentence by analyzing the bodies of text data by using statistical and probabilistic approaches. Internally the model is trained with the training dataset and the classifier file which used to identify the class of the word. The trained model is trained with the help of BERT. In the preprocess step of BERT it uses that every embedding contains three types of embedding namely positional, segment and token embedding. BERT learns and uses positional embedding to express the position of words in a sentence. Token embedding is the embedding learned for the specific token from the WordPiece token vocabulary. For a given token, its input representation is constructed by summing the corresponding token, segment, and position embedding. With the highest probability the relevant class type is identified and with the help of translation dictionary the respective Sinhala definition mapped and it will send back to the browser plugin of which then display the correct definition with the use of a popup box.

3.5 Translation Dictionary

According to “EnSiTip” it created “Sinhala/English translation lexicon” which is having list of Sinhala word with their English meaning (See Figure 3.3).

Figure 3.3: Collection of Sinhala to English meanings

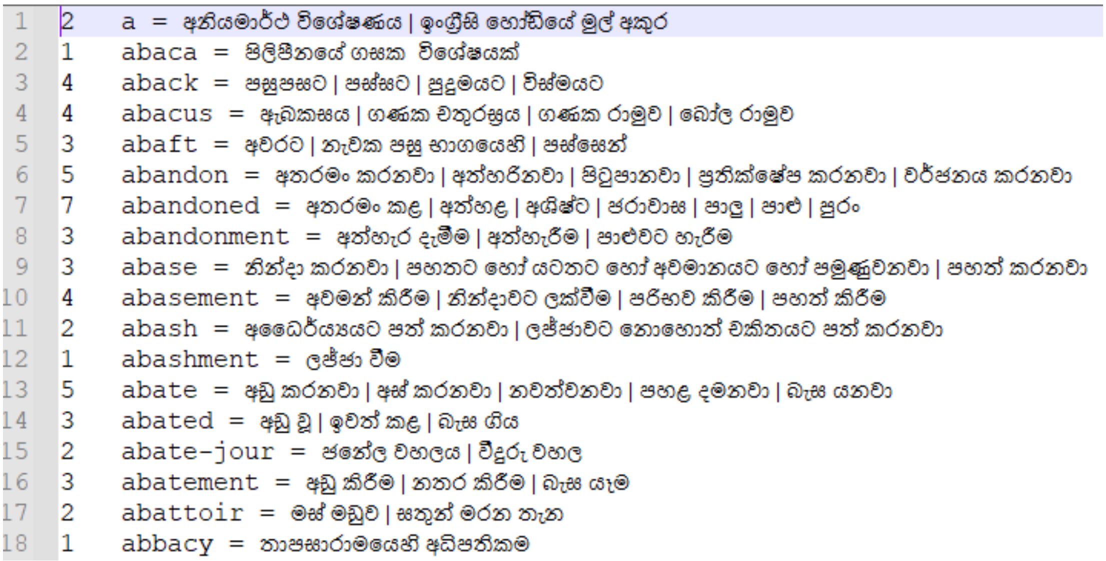


```
43  දැඩි=adamant | bad | bitter | cragged | craggy | deadl
37  මෝඩ=absurd | apish | apishness | asinine | blinking
37  අප්‍රසන්න=awful | bad | beastly | bitter | blatant | com
36  දරුණු=acute | appalling | arrant | brutal | brutish |
33  නියම කරනවා=adjudicate | allot | appoint | assign | as
30  බාධා කරනවා=baffle | balk | bar | barge | baulk | count
29  සමූහය=accumulation | afflux | aggregation | assembly
29  යෝග්‍ය=advisable | applicable | apposite | appropriate
29  නියමය=appointment | assignment | canon | clause | dec
28  සාමාන්‍ය=average | banal | characterless | common | fam
28  රැවටිල්ල=affectation | blind | cheat | cheating | chica
28  මෝඩයා=ass | buffoon | clodhopper | coot | dolt | don
28  ප්‍රධාන=arch | basic | cardinal | central | dominant |
```

In the above Figure the first index indicate the number of definitions for the Sinhala word “දැඩි”. Each different English word which has the same definition “දැඩි” are separated with “|” symbol. This translation dictionary was converted in a way which can be useful to the proposed system. Then created a translation dictionary which having index as number of English words for that English word, English word and the collection of Sinhala definitions relevant to each English word. Then the below translation dictionary was developed. There are more than 36000 English words are listed there (See Figure 3.3). Entries will be validated and cleaned manually. Finally filter and sort according to the alphabet order by making the English word as an index.

According to the below Figure 3.4 English word “abashment” one definition in Sinhala while “abandon” contains five Sinhala definitions. According to the content of the sentence the relevant Sinhala word may use. These indexes help to identify the number of classes which need to specify when building the training model.

Figure 3.4: Collection of English word with Sinhala meanings



```
2    a = අනියමාර්ථ විශේෂණය | ඉංග්‍රීසි හෝඩියේ මුල් අකුර
1    abaca = පිලිපීනයේ ගසක විශේෂයක්
4    aback = පසුපසට | පස්සට | පුදුමයට | විස්මයට
4    abacus = ඇබකසය | ගණක චතුරශ්‍රය | ගණක රාමුව | බෝල රාමුව
3    abaft = අවරට | නැවක පසු භාගයෙහි | පස්සෙන්
5    abandon = අතරමං කරනවා | අත්හරිනවා | පිටුපානවා | ප්‍රතික්ෂේප කරනවා | වර්ජනය කරනවා
7    abandoned = අතරමං කළ | අත්හළ | අශිෂ්ට | ජරාවාස | පාලු | පාළු | පුරං
3    abandonment = අත්හැර දැමීම | අත්හැරීම | පාළුවට හැරීම
3    abase = නින්දා කරනවා | පහතට හෝ යටතට හෝ අවමානයට හෝ පමුණුවනවා | පහත් කරනවා
4    abasement = අවමන් කිරීම | නින්දාවට ලක්වීම | පරිභව කිරීම | පහත් කිරීම
2    abash = අධෛර්ය්‍යයට පත් කරනවා | ලජ්ජාවට නොහොත් චකිතයට පත් කරනවා
1    abashment = ලජ්ජා වීම
5    abate = අඩු කරනවා | අස් කරනවා | නවත්වනවා | පහළ දමනවා | බැස යනවා
3    abated = අඩු වූ | ඉවත් කළ | බැස ගිය
2    abate-jour = ජනේල වහලය | වීදුරු වහල
3    abatement = අඩු කිරීම | නතර කිරීම | බැස යෑම
2    abattoir = මස් මඩුව | සතුන් මරන තැන
1    abbacy = තාපසාරාමයෙහි අධිපතිකම
```

3.6 Language Model

Language model is attempt to characterizes, capture and exploit regularities in natural language. Determine the probability of a word in a sentence by analyzing the bodies of text data by using statistical and probabilistic approaches. This is more specifically called as statistical language models. Large amount of words are used to automatically determine the model's parameter in this statistical language models.

3.7 Training Model Creation

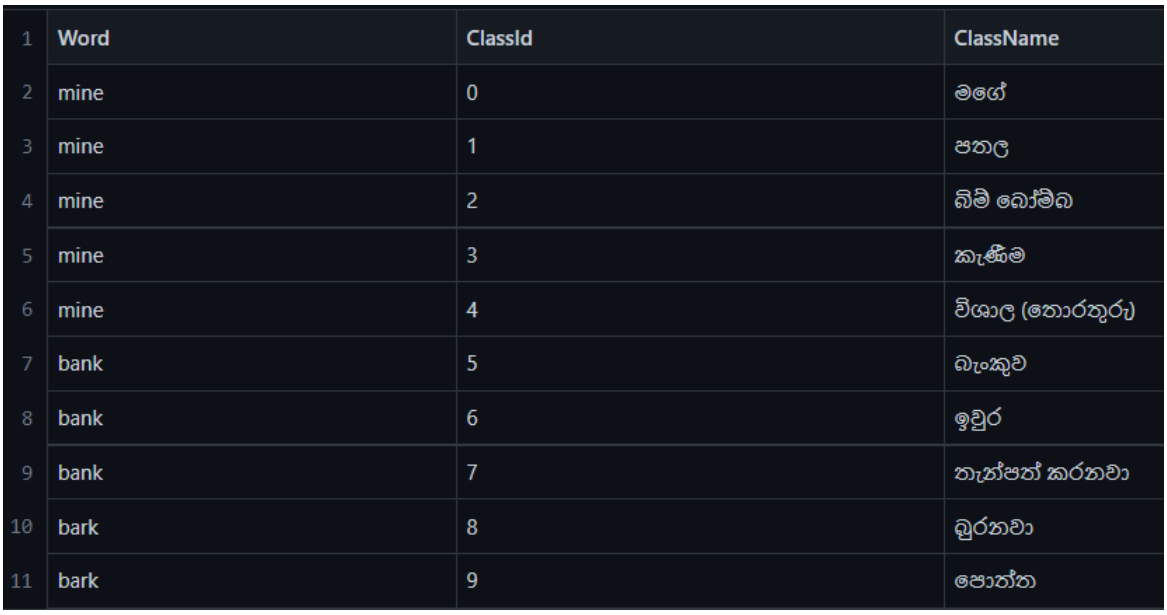
Most of the implementations described in the literature review are purely based on the NLP. Therefore no training dataset is available for this scope to train the model. According to the approach that use for this research, it is mandatory to create a proper datasets for specific English words which have several Sinhala meanings. Training dataset needs to be carefully categorized into relevant classes which make the subclass of specific word. In English Language some words have several different meaning in the sentences which described in the introduction chapter and those words are called homonyms. Different homonyms contain different number of classes. Therefore apart from the training model it needs another model to have all the different classes names with respective the specific word we train. In the test model or in the final output user can select different word of the sentence therefore the specific word also needs to be implemented as dynamic value.

3.7.1 Initial Classifier

For a sentence it contains several numbers of words. Each word of the sentence are belongs to the homonyms. But when considering the web page it may contain more homonyms.

Therefore in this study, first take one homonym as a user selected word and identify the number of classes or the number of meanings that word contains which will be used to train the classifier. The created dataset consist with five main classes (see Figure 3.5). This collection of dataset will be refereed as initial dataset in upcoming sections.

Figure 3.5: Different classes for specific words.



Word	ClassId	ClassName
mine	0	මගේ
mine	1	පතල
mine	2	බිම් බෝම්බ
mine	3	කැණීම
mine	4	විශාල (තොරතුරු)
bank	5	බැංකුව
bank	6	ඉවුර
bank	7	තැන්පත් කරනවා
bark	8	බුරනවා
bark	9	පොත්ත

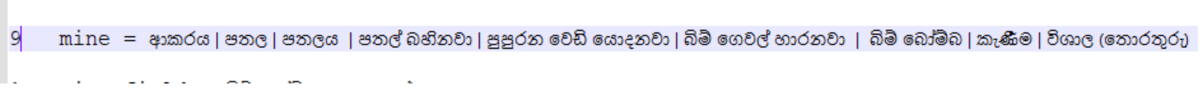
According to the above dataset it mainly contains three specific words they are "Bank", "Bark" and "Mine". Each word assigned with unique "classIds" with unique translations that word contains in the field "ClassName". For example: the word "bank" has two different definitions in the "ClassName" field and each has unique number assigned to "ClassId" field. These "ClassName" are used to display and evaluate the system.

The "ClassId" staring with the "0" it contains number of "ClassId". Consider the word "Mine" for building the initial dataset. These meanings are classified according to the Oxford dictionary.

The author believes it is best and correct approach of classification needed as the output of the selected word. Since human involved in classification of the dataset that use for train and test the model, it can evaluate the approach in a way that output results against the human classified accurate results.

When selecting a homonym as described in the introduction it may contains many different definitions. As mentioned in the above, after building the translation dictionary can identify the number of classes which can be map. But if we carefully look at the created translation dictionary it contains several meanings which complex the implementation (See Figure 3.5).

Figure 3.6: Different meaning for the word "Mine"



```
mine = ආකරය | පතල | පතලය  | පතල් බහිනවා | පුපුරන වෙඩි යොදනවා | බිම් ගෙවල් හාරනවා  |  බිම් බෝම්බ | කැණීම | විශාල (තොරතුරු)
```

Therefore for the simplicity and readability few factors have been taken into consideration. These factors will be described in below.

Factor 1: Remove some of the unnecessary meanings.

According to the above Figure 3.5, Word “Mine” considering the meaning (an excavation in the earth for extracting coal or other minerals) has several similar meanings in Sinhala Language.

Table 3.1: Different definitions in Sinhala to the word "Mine"

Word	Definition in Sinhala Language
Mine	ආකරය, පතල, පතලය, පතල් බහිනවා etc...

Therefore remove the unnecessary similar meanings from the translation dictionary.

Factor 2: Consider only one selected Sinhala definition

For example, the word “Mine” considering the meaning (an excavation in the earth for extracting coal or other minerals) it will only shows the Sinhala word “පතල”.

Table 3.2: One definition highlighted for the word "Mine"

Word	Definition in Sinhala Language
Mine	පතල

This will improve the readability of the user rather than having several similar definitions like showed in Table 3.1.

Factor 3: Focus areas of the words

This model mostly focuses on the nouns, pronouns and verbs. Therefore it is not consider about the tense of the sentences.

Table 3.3: Different definitions of Word "Run" based on tense

Sentence	**Tense**	**Word**	**Definition in Sinhala**
I usually run every day.	Present Simple	Run	දුවනවා
I have never run professionally.	Present Perfect Simple	Run	දුවලා
I think I'll have to run for the bus.	Future Simple	Run	දුවන්න වෙනවා

Because of the complexity, author is not considered many translations for verbs like “Run” based on the tense, therefore when user select the word “Run” it will show the Sinhala definition දුවනවා (See Table 3.4). Author considers improving the system with tenses in the future implementation.

Table 3.4: Selected definition for word "Run"

Word	Definition in Sinhala
Run	දුවනවා.

3.7.2 Initial Training Model

For Sinhala Language it is difficult to find training model for this purpose. Therefore with the help of online dictionaries and resources build the initial training model. As an initial training model, created collection of 120 sentences for using the word “mine” in the sentence which include the different definitions as listed in the Figure 16. yourDictionary is one of the greatest resource to get collection of sentences when building the training model for this kind of scenario.

After collecting these sentences, carefully identifies the meaning of the word “mine” in each sentence. These sentences then map with the “classId” according to the “class name” in Figure 3.3. It contains sentences with the following distribution: 37 are referring to the pronoun (Type 0), 59 are a form of the “පතල” noun (Type 1) and 10 are referring to the “බිම් බෝම්බ” noun (Type 2), 10 are referring to “mine” (verb) (Type 3) and lastly 4 are referring to idiom (Type 4).

In this training model it needs another additional field. That is the “Word” which needs to specify from which word in this sentence it needs to train the sentence. The initial training model will illustrate below (See Table 3.5).

Table 3.5: Initial training model

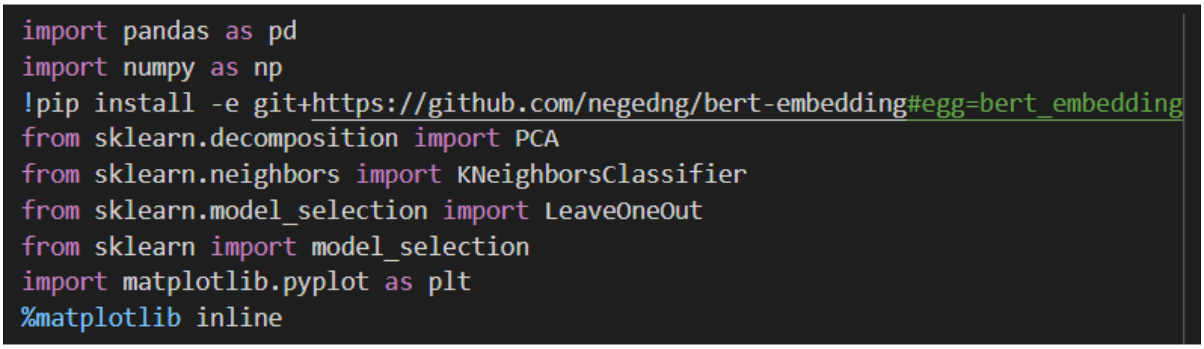
Sentence	Type	Word
"It was safer to leave him in the mine."	1	"mine"
"It was a gold mine, wasn't it?",	1	"mine"
"So I say the horses and chickens are mine and Alex says the other animals are his."	0	"mine"
"You may bring mine with you.",	0	"mine"
"You see, when you die, you have your heaven and I have mine.",	0	"mine"

3.8 Implementation Process

These files saved as comma separated value (CSV) files with the UTF-8 encoding which then easily readable from the Python language. Most of the machine-learning applications are use

Python language to train the model because of the powerfulness and the efficiency. Therefore this system is implemented using the Python language. In order to train the model first need to import the relevant Python libraries (See Figure 3.6).

Figure 3.7: Installed and imported Python libraries

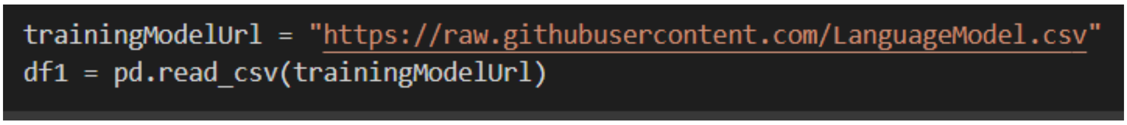


```
import pandas as pd
import numpy as np
!pip install -e git+https://github.com/negedng/bert-embedding#egg=bert_embedding
from sklearn.decomposition import PCA
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import LeaveOneOut
from sklearn import model_selection
import matplotlib.pyplot as plt
%matplotlib inline
```

"pandas" and "numpy" are some of the basic Python libraries which useful when developing the Python applications. The libraries are capable of manipulating high dimensional data and analyses while offering data structures and operations in order to manipulate numerical values. As discussed in above chapters the place of a word in the sentence need to be figure out in order to get the correct definition of that word. To train the sentences in the language model developed earlier for specific word, it uses BERT which considers the relationships by statistical approach to predict the correct definition. Therefore this application uses BERT then it will consider both left and right context of the word before predicting the definition. Scikit-learn (Sklearn) libraries are used for this system because it contains lot of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction. For this system it requires to classify the sentence to correct classes and "matplotlib" library uses for visualize the training model.

Once the required libraries installed then needs to input the training model to train. Created CSV files are easily read from the application when they available in public repository like GIT. These files then read the following Python codes (See Figure 3.7) and store it in a variable which later part of the application then needs to access.

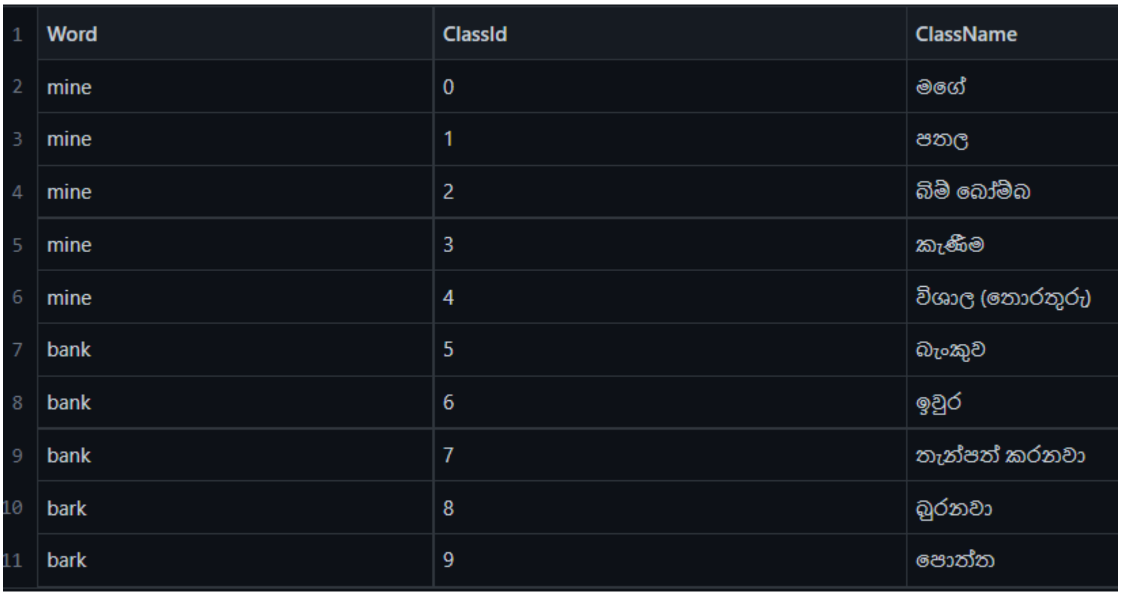
Figure 3.8: Read training model



```
trainingModelUrl = "https://raw.githubusercontent.com/LanguageModel.csv"
df1 = pd.read_csv(trainingModelUrl)
```

Using the "print" command it will display the content of the variable "df1" which it read from the CSV file. The file contains mainly three columns and 120 sentences with the class and the training word respectively (See Figure 3.7).

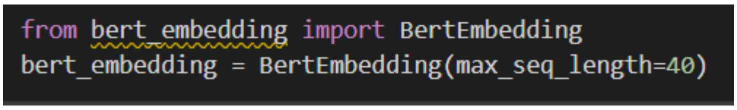
Figure 3.9: Content of the training model



Word	ClassId	ClassName
mine	0	මගේ
mine	1	පතල
mine	2	බිම් බෝම්බ
mine	3	කැණීම
mine	4	විශාල (තොරතුරු)
bank	5	බැංකුව
bank	6	ඉවුර
bank	7	තැන්පත් කරනවා
bark	8	බුරනවා
bark	9	පොත්ත

Then import the "BertEmbedding". In the constructor of the BertEmbedding can specify the max sequence length which denotes the target length of our encodings (See Figure 3.9).

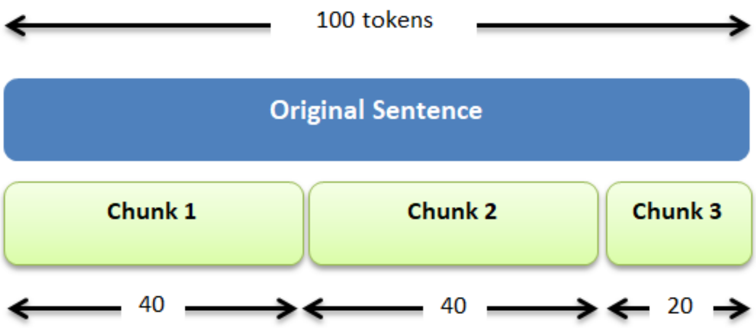
Figure 3.10: Import BERT model



```
from bert_embedding import BertEmbedding
bert_embedding = BertEmbedding(max_seq_length=40)
```

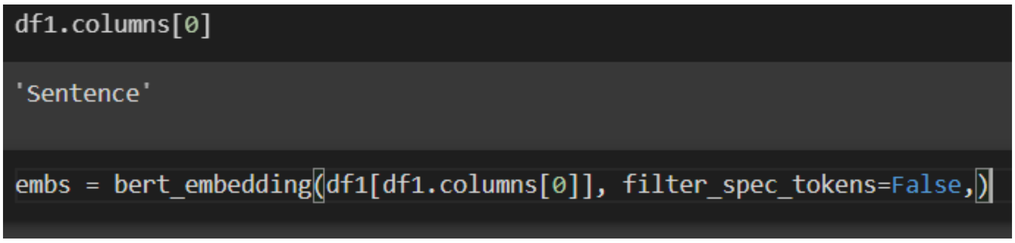
For large piece of text max_seq_length property is need to break the text into small chunks (See Figure 3.10).

Figure 3.11: Chunking mechanism in BERT



According to the Figure 3.9 it illustrates that the model contains 3 columns. Those are "Sentence", "Type" and "Word". Using the column name "Sentence", get all the values of "Sentences" and feed them to the BERT model to tokenize the sentences.

Figure 3.12: Embedding the sentences



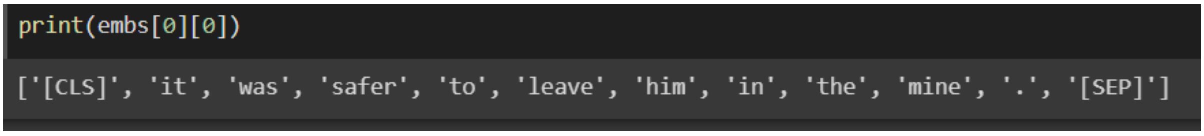
```
df1.columns[0]
```

```
'Sentence'
```

```
embs = bert_embedding(df1[df1.columns[0]], filter_spec_tokens=False,)
```

The output of the above coding (See Figure 3.12) will display the list of tokens, and tokens embedding.

Figure 3.13: Tokens and token embedding for the sentence



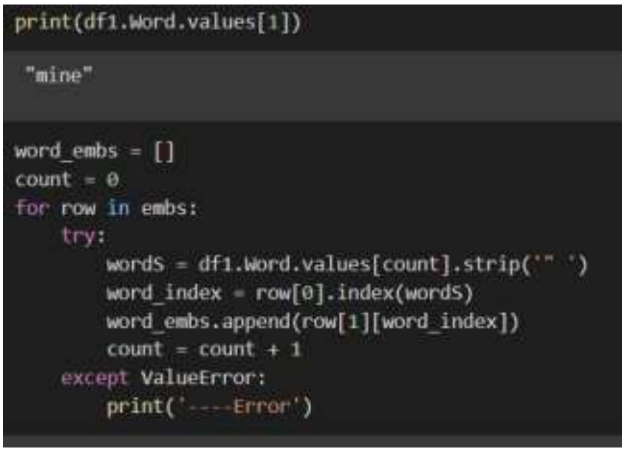
```
print(embs[0][0])
```

```
['[CLS]', 'it', 'was', 'safer', 'to', 'leave', 'him', 'in', 'the', 'mine', '.', '[SEP]']
```

When training large sentences the model itself needs to figure out the sentence starting point and the ending point. Each of the sentences in the CSV file was tokenized by appending the [CLS] and [SEP] tokens. [CLS] indicates the starting point of the sentence and which will appears at the start of every sentence and [SEP] indicates the separate which appears at the ending of each sentences. These tokenized sentences then need to train using the specific word by referring the "Word" column of training CSV file. Using the variable we used to store the file content can retrieve the "Word" column value in each sentence (See Figure 3.13).

Using that word then create the array of tokenized by identifying the index of the specific word in the sentence. Then it the model only the embedding for the 'mine word's token.

Figure 3.14: Train the model with specific word



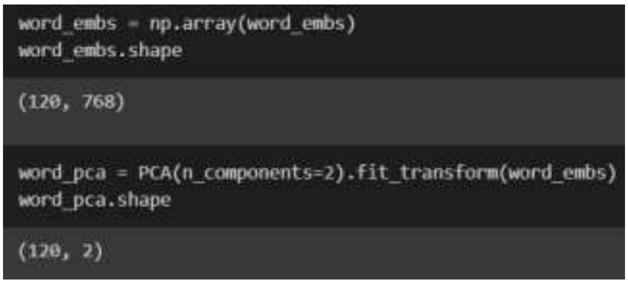
```
print(df1.Word.values[1])
```

```
"mine"
```

```
word_embs = []
count = 0
for row in embs:
    try:
        wordS = df1.Word.values[count].strip('" ')
        word_index = row[0].index(wordS)
        word_embs.append(row[1][word_index])
        count = count + 1
    except ValueError:
        print('----Error')
```

Convert the generated embedding to an array and reshape it using the PCA (See Figure 3.14).

Figure 3.15: Reduce the dimensions using PCA



```
word_embs = np.array(word_embs)
word_embs.shape

(120, 768)

word_pca = PCA(n_components=2).fit_transform(word_embs)
word_pca.shape

(120, 2)
```

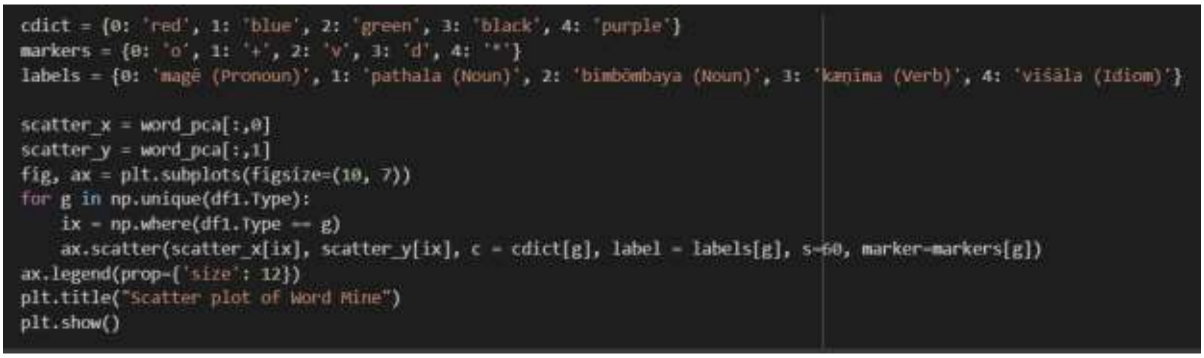
Principal Component Analysis (PCA) is an orthogonal transformation that use in the model to reduce the dimension of the vectors. Using the BERT base uncased model use the last hidden layer which generates 768 size vectors for every word. That is huge number and with the support of PCA it can project 768 dimension vectors to a 2 dimension form where the human can easily visualize the data. PCA keeps the maximum possible variance which means the projection loses information when reducing 738 dimensions to 2 dimensions; therefore it might keep enough variance that helps to identify the classes on the plot.

Figure 3.16: Maximum variant for the first principle component

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w}\right)^2 \right\}$$

Once it reduce to 2 dimension the distribution can be plot easily using the below Python code (See Figure 3.16).

Figure 3.17: Plot the training model

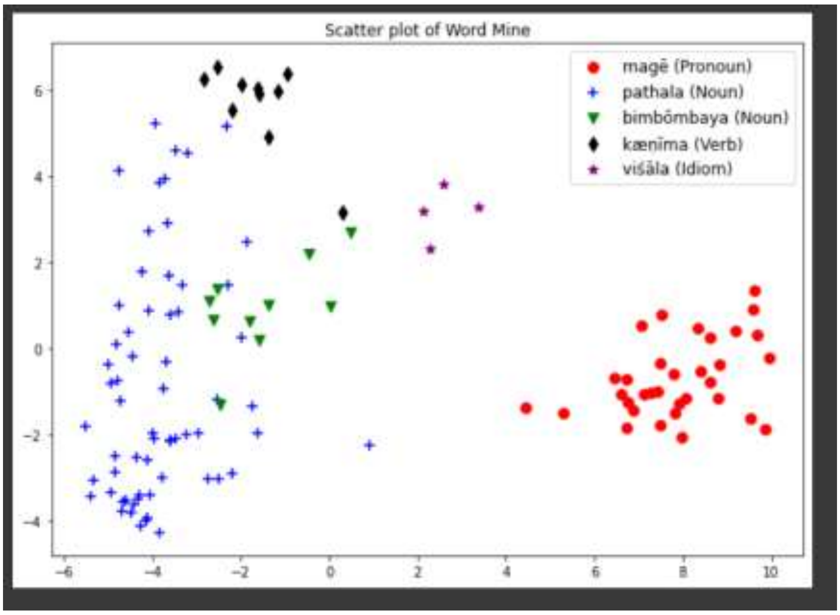


```
cdict = {0: 'red', 1: 'blue', 2: 'green', 3: 'black', 4: 'purple'}
markers = {0: 'o', 1: '+', 2: 'v', 3: 'd', 4: '*'}
labels = {0: 'magē (Pronoun)', 1: 'pathala (Noun)', 2: 'bimbōmbaya (Noun)', 3: 'kæṇima (Verb)', 4: 'viśāla (Idiom)'}

scatter_x = word_pca[:,0]
scatter_y = word_pca[:,1]
fig, ax = plt.subplots(figsize=(10, 7))
for g in np.unique(df1.Type):
    ix = np.where(df1.Type == g)
    ax.scatter(scatter_x[ix], scatter_y[ix], c = cdict[g], label = labels[g], s=60, marker=markers[g])
ax.legend(prop={'size': 12})
plt.title("Scatter plot of Word Mine")
plt.show()
```

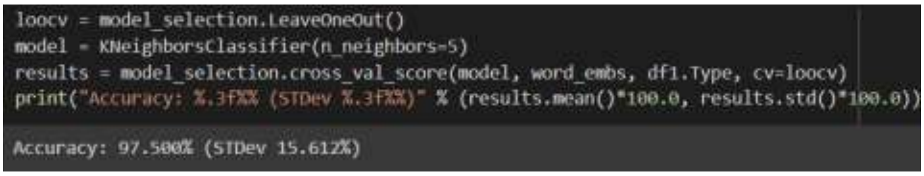
According to the initial classifier build for the word “mine” contains five different classes. Therefore the number of makers and class labels with colors can initialize. The Figure 3.17 shows the result of the projection using the first two principal components. The classes are manually annotated types according to the classifier. It clearly illustrates the type pronoun and type idiom easily separated from the others.

Figure 3.18: Scatter plot of the word "Mine"



Since the dataset is small, k-NN classifier can be used because it uses the k closest samples to predict the class of a new sample. K-NN selects the most represented class from the neighborhood. Leave One Out Cross-Validation (LOOCV) is use for validating the model. As name implies it take 1 sample out of 120 samples and the remaining 119 which will validate using the single sample. Accuracy can be generated using the results obtained (See Figure 3.18).

Figure 3.19: Accuracy of the training model



```
loocv = model_selection.LeaveOneOut()
model = KNeighborsClassifier(n_neighbors=5)
results = model_selection.cross_val_score(model, word_embs, df1.Type, cv=loocv)
print("Accuracy: %.3f%% (STDev %.3f%%)" % (results.mean()*100.0, results.std()*100.0))

Accuracy: 97.500% (STDev 15.612%)
```

3.9 Test Model

Test model can create as same for the training model by having new sentences which are not included in the training model and most importantly without specifying the class name for each sentence. Indicating the “Sentence” and “Word” as column names, create a simple test model with six sentences (See Table 3.6).

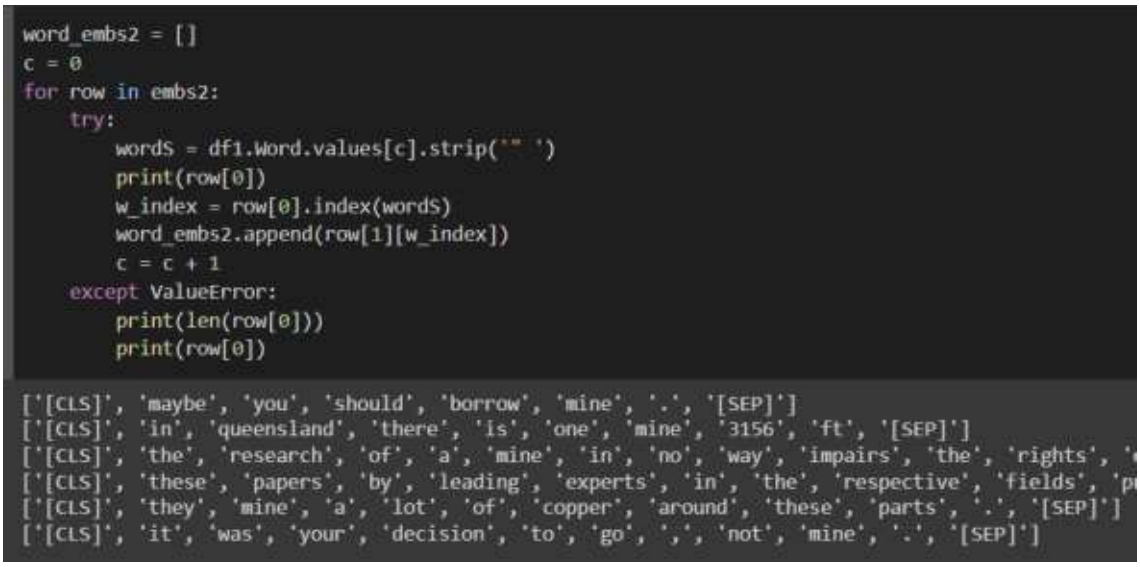
Table 3.6: Test Model

Sentence	Word
"Maybe you should borrow mine.",	"mine"
"In Queensland there is one mine 3156 ft"	"mine"
"The research of a mine in no way impairs the rights of ownership of the land in	"mine"

which the mine is located."	
"These papers by leading experts in the respective fields provide a mine of information that will be referred to for some time to come"	"mine"
"They mine a lot of copper around these parts."	"mine"
"It was your decision to go, not mine."	"mine"

The same process of the training model can use to predict the class name of each sentence. Each tokenize sentence can display as in the Figure 3.19.

Figure 3.20: Tokenized sentences for the test model

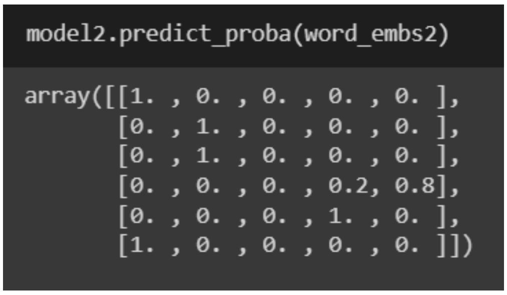


```
word_embs2 = []
c = 0
for row in embs2:
    try:
        wordS = df1.Word.values[c].strip('" ')
        print(row[0])
        w_index = row[0].index(wordS)
        word_embs2.append(row[1][w_index])
        c = c + 1
    except ValueError:
        print(len(row[0]))
        print(row[0])
```

```
['[CLS]', 'maybe', 'you', 'should', 'borrow', 'mine', '.', '[SEP]']
['[CLS]', 'in', 'queensland', 'there', 'is', 'one', 'mine', '3156', 'ft', '[SEP]']
['[CLS]', 'the', 'research', 'of', 'a', 'mine', 'in', 'no', 'way', 'impairs', 'the', 'rights', '
['[CLS]', 'these', 'papers', 'by', 'leading', 'experts', 'in', 'the', 'respective', 'fields', 'p
['[CLS]', 'they', 'mine', 'a', 'lot', 'of', 'copper', 'around', 'these', 'parts', '.', '[SEP]']
['[CLS]', 'it', 'was', 'your', 'decision', 'to', 'go', ',', 'not', 'mine', '.', '[SEP]']
```

The variable “model2” indicate the test model which needs to display the predicted class of each Sentence. The score of the prediction can be view as below (See Figure 3.20).

Figure 3.21: Sentences showing the probabilities for each class

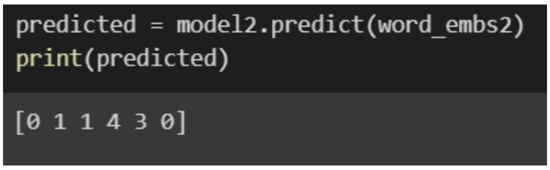


```
model2.predict_proba(word_embs2)
```

```
array([[1. , 0. , 0. , 0. , 0. ],
       [0. , 1. , 0. , 0. , 0. ],
       [0. , 1. , 0. , 0. , 0. ],
       [0. , 0. , 0. , 0.2, 0.8],
       [0. , 0. , 0. , 1. , 0. ],
       [1. , 0. , 0. , 0. , 0. ]])
```

According to the results generate, in the fourth row; the model is 20% predicting that sentence is belongs to “Type 3” and 80% predicting to “Type 4”. Final value is based on the highest probability which is “Type 4”. The predicted class for the sentences in testing model is shown below (See Figure 3.21).

Figure 3.22: Predicted classes for test model



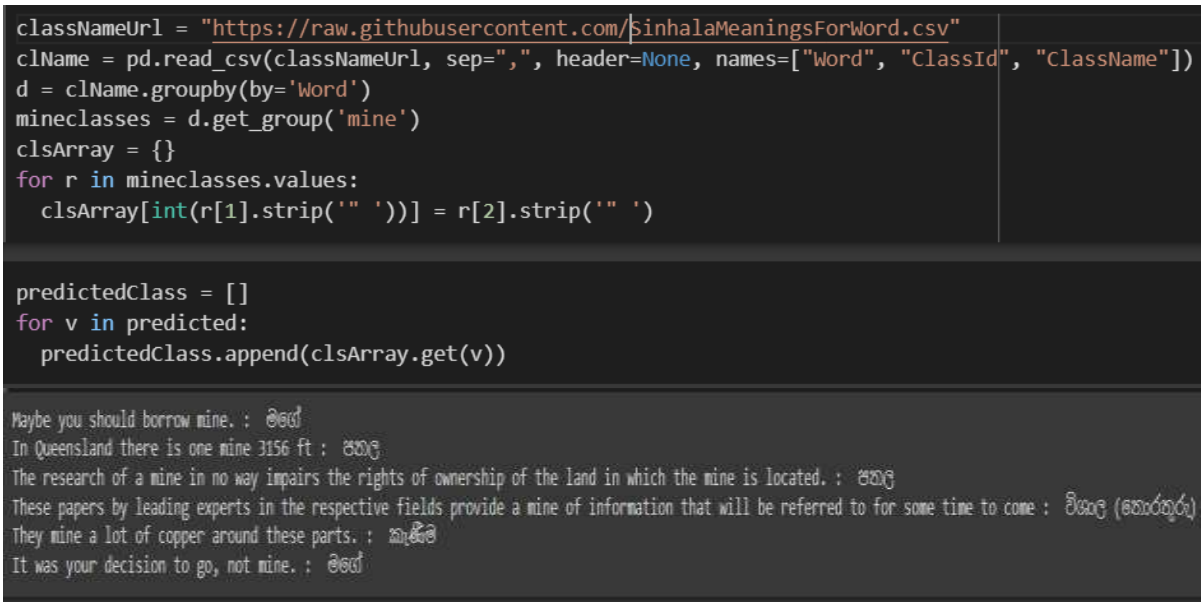
```
predicted = model2.predict(word_embs2)
print(predicted)
```

```
[0 1 1 4 3 0]
```

Using the classifier file developed above can display the class name which is more readable.

The results of the test model display as below (See Figure 3.23). According to the sentences in the test model all the six sentences are correctly classify to its specific class.

Figure 3.23: Output of the test model



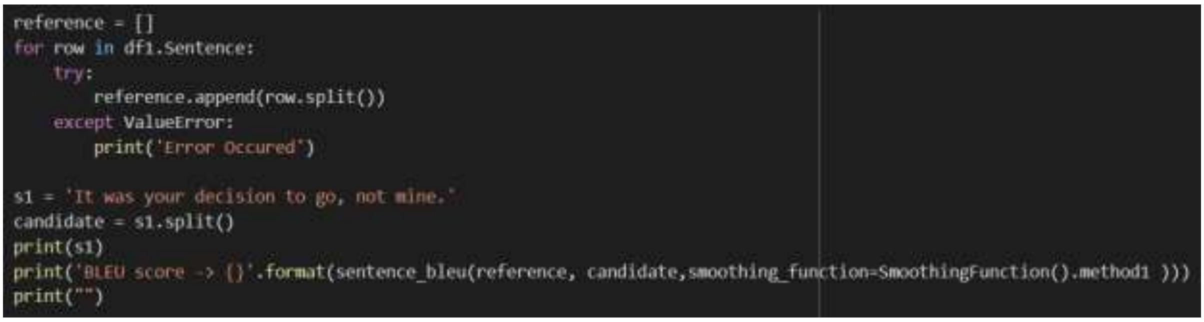
```
classNameUrl = "https://raw.githubusercontent.com/SinhalaMeaningsForWord.csv"
clName = pd.read_csv(classNameUrl, sep=",", header=None, names=["Word", "ClassId", "ClassName"])
d = clName.groupby(by='Word')
mineclasses = d.get_group('mine')
clsArray = {}
for r in mineclasses.values:
  clsArray[int(r[1].strip('" '))] = r[2].strip('" ')
```

```
predictedClass = []
for v in predicted:
  predictedClass.append(clsArray.get(v))
```

```
Maybe you should borrow mine. :  මගේ
In Queensland there is one mine 3156 ft :  පතල
The research of a mine in no way impairs the rights of ownership of the land in which the mine is located. :  පතල
These papers by leading experts in the respective fields provide a mine of information that will be referred to for some time to come :  ඕපල (තොරතුරු)
They mine a lot of copper around these parts. :  කැණීම
It was your decision to go, not mine. :  මගේ
```

BLEU score can be measured for test model based while having the training model as reference.

Figure 3.24: Code to display BLEU score for the sentence

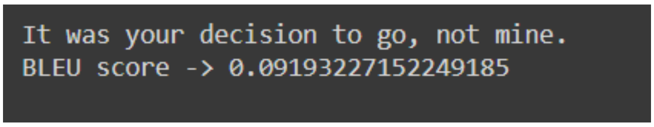


```
reference = []
for row in df1.Sentence:
    try:
        reference.append(row.split())
    except ValueError:
        print('Error Occured')

s1 = 'It was your decision to go, not mine.'
candidate = s1.split()
print(s1)
print('BLEU score -> {}'.format(sentence_bleu(reference, candidate,smoothing_function=SmoothingFunction().method1 )))
print("")
```

The BLEU score indicating is 0.09193 on Figure 38 which illustrates that the exact sentence is not included in the training model and the test model correctly identifies the class accordingly.

Figure 3.25: BLEU score for the sentence



```
It was your decision to go, not mine.
BLEU score -> 0.09193227152249185
```

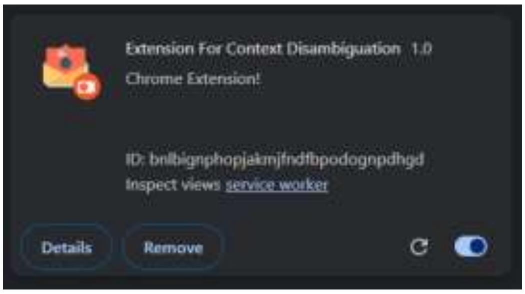
As discussed above the initial training model is created with the word “mine” and the training model can add sentences which train for another set of words. Same as for the initial classifier can add another set of words by identifying different classes of that word. Application can train the model and implement it as a browser plugin. Then the plugin will display the correct Sinhala definition when user mouse-hover to the English word in web content.

3.10 Chrome Plugin

Implementing a Chrome plugin entails the development of a browser extension aimed at augmenting the functionality of the Google Chrome web browser. These plugins, commonly referred to as Chrome extensions, are crafted utilizing web technologies such as HTML, CSS, and JavaScript.

The final deliverable of the proposed system is a browser plugin. This implemented plugin comprises several key files, including the background JavaScript file, manifest JSON file, and Python scripts. These files collectively capture mouse-hovered text and relevant sentences from web pages, subsequently presenting the output as a tooltip. The plugin's functionality involves utilizing a pre-trained model or, due to the extensive lexicon involved in this research, identifying and incorporating 15 ambiguous words. For all other non-ambiguous words, the plugin retrieves a general Sinhala translation from the dictionary. However, if a word lacks a corresponding Sinhala translation, the plugin returns "Not found" as the output.

Figure 3.26: Chrome Extension



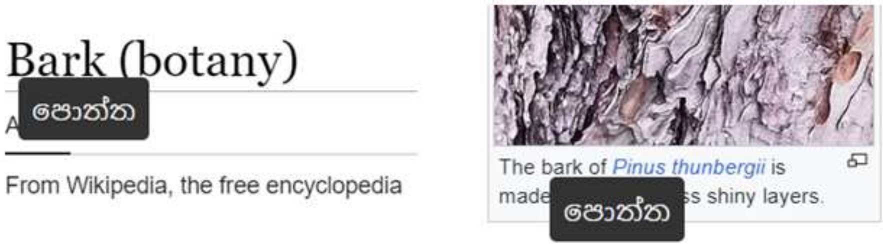
Extension For Context Disambiguation 1.0
Chrome Extension!

ID: bnlbignphopjakmjfndfbpodognpdhgd
Inspect views service worker

The preceding figure depicts the interface when the extension is activated within the Chrome browser.

The below figures 41 and 42 shows the two different translations for the English word “Bark” correctly identified by the developed plugin based on its context.

Figure 3.27: Showing Sinhala translation for word Bark



Bark (botany)

පොත්ත

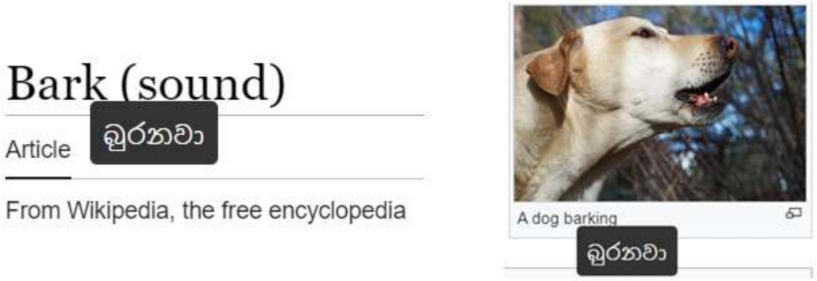
A

From Wikipedia, the free encyclopedia

The bark of *Pinus thunbergii* is made ss shiny layers.

පොත්ත

Figure 3.28: Showing Sinhala translation for word Bark



Bark (sound)

බුරනවා

Article

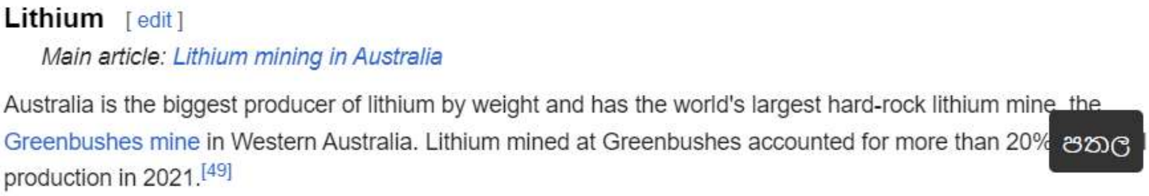
From Wikipedia, the free encyclopedia

A dog barking

බුරනවා

The below figures 3.27 and 3.28 shows the two different translations for the English word “Mine” correctly identified by the developed plugin based on its context.

Figure 3.29: Showing Sinhala translation for word Mine



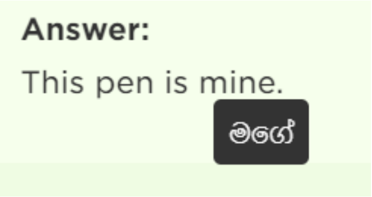
Lithium [edit]

Main article: Lithium mining in Australia

Australia is the biggest producer of lithium by weight and has the world's largest hard-rock lithium mine, the Greenbushes mine in Western Australia. Lithium mined at Greenbushes accounted for more than 20% production in 2021.[49]

පතල

Figure 3.30: Showing Sinhala translation for word Mine



Answer:

This pen is mine.

මගේ

CHAPTER 4
EVALUATION AND RESULTS

4.1 Introduction

This section provides a summary of quantitative and qualitative assessments, drawing from feedback obtained from diverse users. These users comprise domain experts, academic scholars, researchers, and the general community. The feedback spans various facets of the research and prototype, encompassing aspects like the uniqueness of the concept and the caliber of the final product.

4.2 Evaluation Methodology and Approach

The evaluation process serves as a means to assess the effectiveness of a project. In the research study entitled "Context-Sensitive Translation of English Words and Phrases to Sinhala in Web Content," the primary focus is on proposing an innovative framework for ensuring accurate Sinhala translations of English terms when users hover over them. Therefore, the author has opted to utilize both qualitative and quantitative methodologies to evaluate the project's success. Qualitative data analysis will involve the application of deductive thematic techniques. In addition to qualitative assessments, quantitative feedback will be gathered during the evaluation phase and subsequently analyzed and presented in the dedicated quantitative evaluation section. Distribution of questionnaires has been undertaken to solicit feedback from users.

4.3 Evaluation criteria

4.3.1 Qualitative Evaluation

The subsequent evaluation criteria have been established in accordance with the principles outlined for the evaluation process. In order to understand how well the potential user group will engage with the implemented solution, it's important to assess their understanding of it. Qualitative analysis will be conducted based on the criteria listed below, and the form utilized to gather qualitative feedback is provided in the appendix.

Evaluation criteria

Table 4.1: List of evaluation criteria

Criteria	Purpose
Education level / Stream of education / Age / Gender / English literacy level	This evaluation helps determine how familiar users are with the solution and its features. By gauging users' comprehension, author gain valuable insights into whether the solution meets their needs and expectations. This evaluation process is essential for ensuring that the solution is user-friendly and effectively addresses the requirements of the intended user group.
Currently using plugins for translating to Sinhala	To provide insights into user behaviors, preferences, and market demand, thereby informing product development strategies and enhancing user-centric approaches within the realm of translation tools.
Primary browser	Knowledge of preferred browsers informs decisions regarding the implementation of new features/ technologies, enabling author to focus efforts on areas that will have the greatest impact on user satisfaction in future releases.
Correctness of the plugin	To assess the accuracy of the implemented solution.
Satisfaction of the response time of the translated Sinhala word	To evaluate the users expectation and willingness to use this plugin
Satisfaction of the tooltip color / recommendation tooltip color	To evaluate the user experience in order to enhance the user experience
Satisfaction for the implemented Plugin / recommendation this Chrome Plugin to others.	To assess user satisfaction and the likelihood of recommendation to others.

The selection of evaluators will be determined by specific requirements and expertise criteria.

Table 4.2: List of evaluator's categories

Category	Evaluator Type	Requirement	Expertise Level
L1	Domain and Technical Experts	Satisfaction / Correctness / Recommendation / Suggestions for plugin	Graduates who have experience in the area
L2	Beginner Researchers	Correctness / Recommendation / Suggestions for plugin	Graduates and Undergraduate
L3	General Community	Evaluating the overall solution	Technical Fluency to use an application

Qualitative Evaluation Results

The following table presents the details of the evaluators who assessed the system.

Table 4.3: Academic level of evaluators

Id	Academic Level	Category
1	Phd (Reading)	L1
2	Msc (Reading)	L1, L2
3	Msc (Reading)	L1, L2
4	Bsc	L1, L2
5	Bsc	L2
6	Bsc	L2
7	Msc	L1

The feedback received from the survey is outlined below.

Table 4.4: Feedbacks received

Id	Feedback
1	This is a good piece of work. The researcher seems to have put a good effort into building a robust model for WSD in Sinhala. Keep up the excellent work.
2	Great plugin i suggest to be used by everyone
4	It is a useful plugin because as a person I don’t know the meaning of all the English words. So it is very helpful
5	Good work
6	This is useful
7	Good Research

4.3.2 Quantitative Evaluation

The quantitative analysis in this research delineates the model and system's credibility through below quantitative evaluations.

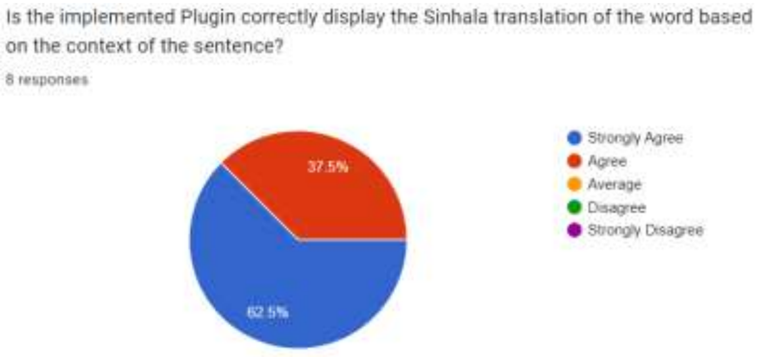
Table 4.5: List of qualitative evaluations

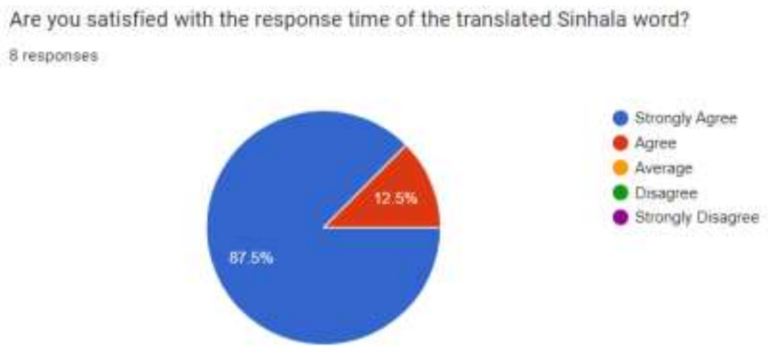
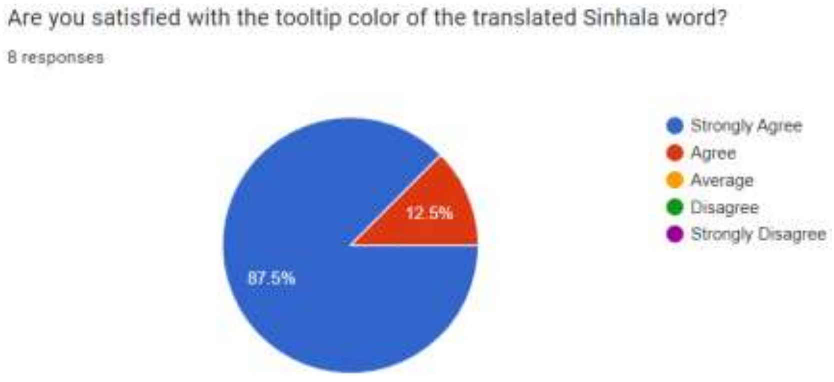
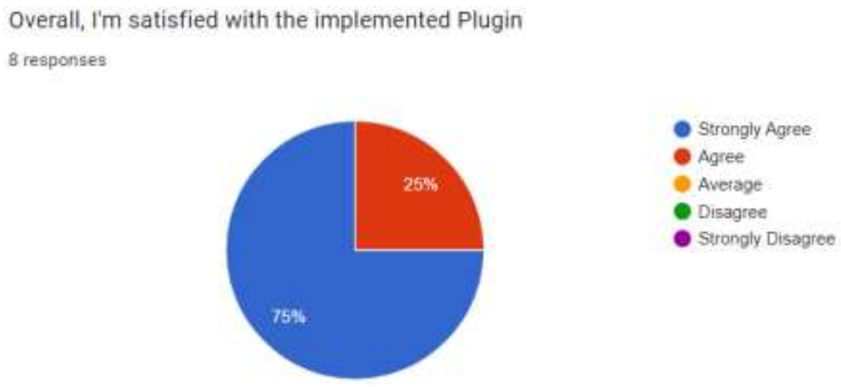
Id	Evaluating Section	Evaluating Technique	Description
1	Correctness of the plugin	Likert scale	Strongly Agree to Strongly Disagree
2	Satisfaction of the response time of the translated Sinhala word	Likert scale	Strongly Agree to Strongly Disagree
3	Satisfaction of the tooltip color	Likert scale	Strongly Agree to Strongly Disagree
4	Satisfaction for the implemented Plugin	Likert scale	Strongly Agree to Strongly Disagree
5	Plugin recommendation to others	Likert scale	Strongly Agree to Strongly Disagree

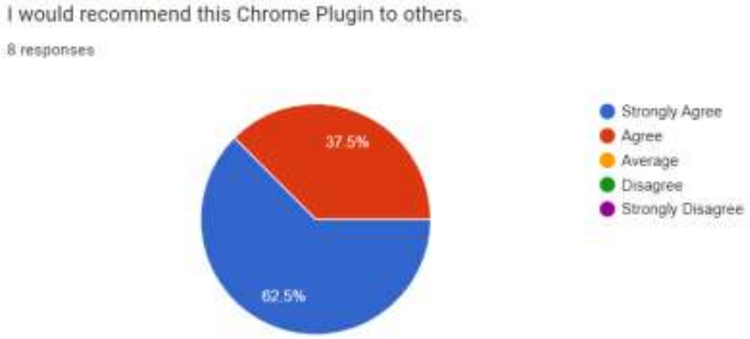
Quantitative Evaluation Results based on the evaluation

Eight evaluators have completed the evaluation form. The summary of the results is presented below.

Table 4.6: Response received for evaluation

<table>
<tr><td colspan="2">Evaluation Criteria : Correctness of the plugin</td></tr>
<tr><td>Evaluation Result:

</td><td>Discussion:
62.5% of the evaluators strongly agree that the plugin correctly display the Sinhala translation for the English words, and 37.5% percentage agree on the same.</td></tr>
<tr><td colspan="2">Review: Users are satisfied with the browser plugin to provide context disambiguate Sinhala translation for English words.</td></tr>
</table>

<table>
<tr><td>Evaluation Result:

Are you satisfied with the response time of the translated Sinhala word?
8 responses
 </td><td>Discussion:
87.5% of the evaluators strongly agree with the response time of the implemented plugin, and 12.5% percentage agrees on the same.</td></tr>
<tr><td colspan="2">Review: Users satisfied with the response time of the plugin, because it displays the Sinhala translation as a tooltip when user mouse hovered to the word.</td></tr>
<tr><td>Evaluation Result:

Are you satisfied with the tooltip color of the translated Sinhala word?
8 responses
 </td><td>Discussion:
87.5% of the evaluators strongly agree with the tooltip color of the implemented plugin, and 12.5% percentage agrees on the same.</td></tr>
<tr><td colspan="2">Review: Users satisfied with the color of the tooltip in the implemented plugin.</td></tr>
<tr><td>Evaluation Result:

Overall, I'm satisfied with the implemented Plugin
8 responses
 </td><td>Discussion:
75% of the evaluators strongly agree with the satisfaction of overall implemented plugin, and 25% percentage agrees on the same.</td></tr>
<tr><td colspan="2">Review: Users satisfied with the overall plugin implemented.</td></tr>
<tr><td>Evaluation Result:</td><td>Discussion:
62.5% of the</td></tr>
</table>

 	evaluators strongly agree that they recommend implemented plugin to others, and 37.5% percentage agree on the same.
Review: Most of the users recommend the plugin to others	

4.3.3 Self Evaluation

Table 4.7: Self evaluation

Criteria	Discussion
Concept of the research	The main aim of the research is to build a browser plugin which support users to get correct Sinhala translation for the mouse hovered English words and increase their literacy level unintentionally.
Scope of the project	The Scope has been discussed under section 1.12.
Solution	The plugin implemented for the Chrome browser which is user friendly. The result of the Sinhala translation word provide when mouse hovered.
Limitations and future enhancements	The implemented plugin is trained for limited set of words which have ambiguity and only provide the result for stem of the words but it gives the other words translations from the dictionary. The training model can increase the words not only for stem but also for other words. Since this project is implemented targeting the Sri Lankans this plugin also can enhance for worldwide users by giving the translation in English.

4.3.4 Numeric Evaluation

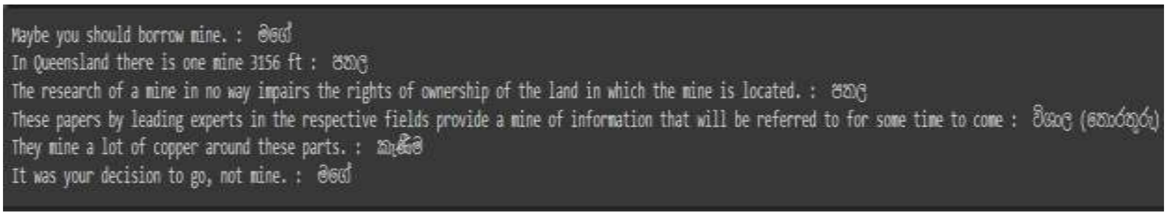
BLEU, or Bilingual Evaluation Understudy, represents a pivotal metric in assessing the fidelity of machine-generated translations against human-crafted reference translations. Through a systematic analysis of n-gram precision and a weighted geometric mean, BLEU quantifies the degree of overlap between candidate translations and reference translations. The

BLEU score is a number between zero (0) and one (1) that measures the similarity of the machine-translated text to a set of high-quality reference translations.

As for the evaluation approach for this system the author created a separate test model contains several sentences specifying an English word which needs to get the Sinhala translation. With the support of BLEU score with the training model and test model, it is capable to identify the sentence in test model is available in the training dataset or not. An accurate and practical model can be beneficial for users to take correct meaning of the English word in Sinhala within milliseconds by mouse-hovering the word. This sample test model is created and validate.

Result of the test model

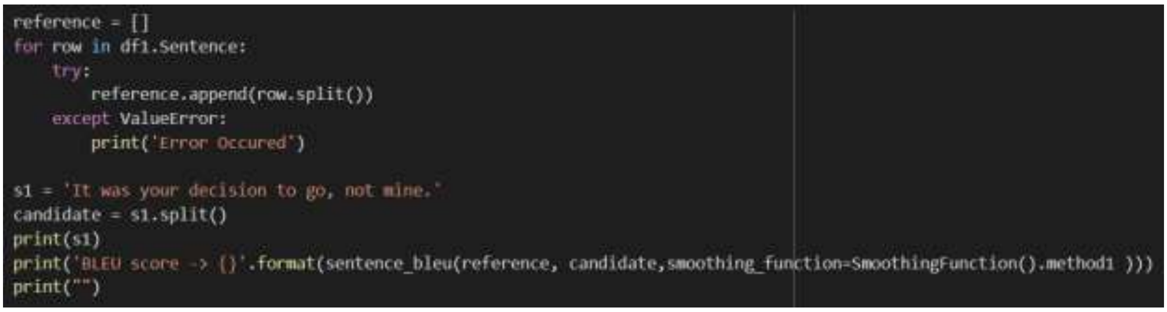
Figure 4.1: Output of the test model



```
Maybe you should borrow mine. :  මගේ
In Queensland there is one mine 3156 ft :  පතල
The research of a mine in no way impairs the rights of ownership of the land in which the mine is located. :  පතල
These papers by leading experts in the respective fields provide a mine of information that will be referred to for some time to come :  විශාල (තොරතුරු)
They mine a lot of copper around these parts. :  කැණීම
It was your decision to go, not mine. :  මගේ
```

BLEU score can be measured for test model based while having the training model as reference.

Figure 4.2: Code to display BLEU score for the sentence

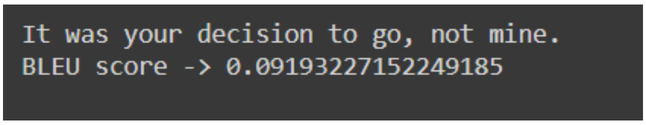


```
reference = []
for row in df1.Sentence:
    try:
        reference.append(row.split())
    except ValueError:
        print('Error Occured')

s1 = 'It was your decision to go, not mine.'
candidate = s1.split()
print(s1)
print('BLEU score -> {}'.format(sentence_bleu(reference, candidate,smoothing_function=SmoothingFunction().method1 )))
print("")
```

The BLEU score indicating is 0.09193 on Figure 48 which illustrates that the exact sentence is not included in the training model and the test model correctly identifies the class accordingly.

Figure 4.3: BLEU score for the sentence



```
It was your decision to go, not mine.
BLEU score -> 0.09193227152249185
```

Accuracy can calculate from the below formula,

Word Accuracy = $\frac{\sum number\ of\ correctly\ identified\ sentences}{Total\ sentences}$ %

Creating a test model with 100 sentences including all the ambiguous words used to train the model gave the accuracy of 0.89 from the above calculation.

4.3.5 Evaluation of Functional Requirements

The evaluation of both functional and non-functional requirements will be conducted according to the criteria outlined below.

Table 4.8: List of functional requirements

Id	Requirement	Priority	Completion Status
1	User should be able to mouse hover the word to get the Sinhala translation	High Priority	Completed
2	Plugin is capable of sending the response based on the selected word	High Priority	Completed
3	Plugin is capable to capture the mouse hovered word correctly.	High Priority	Completed
4	Plugin is capable of capture the sentence based on the mouse hovered word.	High Priority	Completed
5	The response is showing as a tooltip on the mouse hovered word	High Priority	Completed
6	Plugin is capable to provide relevant response for the word which is not in the training model.	High Priority	Completed
Functional Requirement completion rate 100%			

4.3.6 Evaluation of the Research Questions

Table 4.9: Evaluation for research questions

Question	RQ1: What are the major challenges when building a language model to get context sensitive Sinhala translation for a selected English word/phrase in the web content?
Answer	There are many challenges. These challenges encompass the complexity of the Sinhala language, the context sensitivity required for accurate translations, ambiguity in English expressions, limited availability of training data, domain-

	specific terminology, idiomatic expressions, proper noun translation, and computational efficiency.
Question	RQ2: What attributes can build an accurate model?
Answer	This is relies on several key attributes such as quality data, feature selection, feature engineering, model selection, cross-validation, and domain knowledge.
Question	RQ3: What are the limitations of such a language model and how to improve?
Answer	Data bias, context understanding, out of domain data, rare or unseen events are some of the limitations for creating a language model. These can improve by incorporating diverse and representative training data, implementing continual learning techniques allows language models to adapt and improve over time as they encounter new data and scenarios and applying regularization techniques helps prevent over fitting and improves the generalization ability of language models.
Question	RQ4: How can the vast knowledge encoded in the English web be understood by native Sinhala speakers while not depriving them of the opportunity to learn English?
Answer	Implemented solution for browser plugin enhance the readers literacy, language learning platforms, Bilingual education, Cultural integration and promotion of English language resources.
Question	RQ5: What kind of language model would best facilitate this goal?
Answer	Language model with translation capability, contextual understanding, adaptability to user proficiency levels, language learning resources and community engagement and continuous improvement.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

Ultimate goal of this project was to design a system which is capable of providing word sense disambiguation, while translating English word in the web page to correct meaning of the word in Sinhala Language. Once the related work under this domain is studied, the author's attempt has made to translate the English word into Sinhala meaning with the use of training model in machine learning approach. Different technologies and concepts were followed for catering this purpose such as Bidirectional Encoder Representations from Transformers (BERT).

As mention in the above sections of the thesis, the system was developed to recognize thirteen ambiguity (homonyms) English words in order to make the training model which is the core component of the whole process.

The scope of this study is limited to English-based web pages and websites in consideration of prevailing usage trends among Sri Lankans. Statistics indicate a significant portion of the population primarily engages with English web content for daily activities. This study does not encompass non-English user communities. The system design includes a feature where only mouse-hovered English words translate into Sinhala. This approach aims to provide room for readers to learn English, focusing on individual word translations rather than translating entire sentences.

Another important goal was to make the system easy to train. Within the system thirteen ambiguous English words were selected to train the model using 1277 sentences and more English words can be added to the system in the pre training process.

To enhance the accuracy of the system, it is imperative to consider several strategies. Firstly, expanding the training dataset by including more sentences for specific English words would improve the system's ability to discern varied contexts. Concurrently, enlarging the system's vocabulary by integrating additional English words would bolster its overall proficiency. Ensuring proper labeling of new training words with correct class types is essential to prevent overlap with existing classes and maintain data integrity. Moreover, when incorporating new training sentences, adherence to the principle of including the exact match of the training word only once per sentence is crucial. This meticulous approach not only fosters better accuracy but also minimizes ambiguity in the model's learning process, facilitating precise associations between words and their contextual usage. By implementing these measures, the system stands to significantly enhance its accuracy, thereby bolstering its effectiveness and reliability in linguistic interpretation and analysis.

From the users' perspective, minimizing system costs stands as a paramount objective. The hardware and software requirements elucidate this commitment to affordability, as the system does not necessitate specialized or costly devices for operation. Users merely require a basic computer to access the web page and activate the browser plugin, provided they possess a stable network connection for seamless viewing of Sinhala meanings. Additionally, the system's user interface prioritizes simplicity and intuitiveness, requiring users only to hover the cursor over the desired word to retrieve its Sinhala meaning, eliminating the need for additional user actions. This design approach ensures a user-friendly experience, enhancing accessibility while minimizing operational complexities and associated costs.

Extensive research and analysis were conducted to gain a comprehensive understanding of the project domain. The investigation commenced with an exploration of word ambiguity types, which entailed an exhaustive literature review consuming a significant amount of time. Subsequently, attention shifted towards leveraging articles and publications to discern accurate word meanings, employing BERT as the underlying framework. Notably, the author of a pivotal article proposed a methodology for disambiguating single words, serving as a foundation for subsequent endeavors. Various methodologies were systematically explored through multiple iterations, with a focus on adapting machine learning approaches to address ambiguity in Sinhala language contexts. This endeavor involved a thorough examination of prevalent techniques and methodologies within the field, enabling a nuanced understanding of their respective strengths and limitations. Despite the absence of a robust implementation in Sinhala, efforts persisted in training models and assessing techniques pertinent to the domain. Through meticulous study and analysis, a comprehensive comprehension of the methodologies, techniques, and challenges associated with disambiguation in Sinhala language was attained.

Thorough system analysis was conducted, culminating in the development of a robust design for the proposed system. Each component and sub-component of the system underwent meticulous identification, delineating their respective functionalities and interconnections. Transitioning to the implementation stage, a multitude of challenges and issues emerged, necessitating the adoption of diverse techniques to address them effectively. The endeavor demanded considerable effort, dedication, and time investment to forge a stable prototype system capable of reliably providing accurate meanings of words in the Sinhala language. Through persistent iteration and problem-solving, the team succeeded in overcoming hurdles and achieving the desired outcome, exemplifying a commitment to excellence and proficiency in system development.

As outlined previously, the proposed system adopts a machine learning approach for model training. This pre-trained model is meticulously crafted by the author through manual analysis, meticulously scrutinizing the classes assigned to each sentence. However, it is pertinent to note three key considerations that were not factored into the system's development methodology. Firstly, the process involves the removal of unnecessary meanings, restricting the focus to a single selected definition, and limiting the scope to nouns, pronouns, and verbs exclusively. This strategic decision aims to mitigate unnecessary complexities within the system. While pre-trained models for the English language exist, previous research has primarily focused on comparing the similarity and dissimilarity of ambiguous words based on context using cosine values, rather than generating outputs for the correct meaning of words in other relevant languages. This delineation underscores the unique approach undertaken by the proposed system, emphasizing its novel contribution to the field of language processing and disambiguation.

A sample implementation was constructed using Google Colab and subsequently integrated into the local environment to culminate in the final prototype, presented as a browser plugin. The prototype's functionality is constrained by a limited selection of ambiguous English words, primarily due to the inherent challenges associated with generating a pre-trained model encompassing a vast array of English words along with their respective sentences. Furthermore, time constraints imposed limitations on the scope and depth of the implementation process. Despite these constraints, the prototype serves as a testament to the feasibility and potential of the proposed system, offering valuable insights into the practical implications of its design and functionality within a real-world context.

The process of system training and evaluation underwent thorough scrutiny to assess both accuracy and usability. This comprehensive evaluation is delineated in detail within the Evaluation section. Throughout this phase, meticulous attention was devoted to validating the system's performance across various metrics and criteria. By systematically analyzing the results of the evaluation process, insights were gleaned regarding the system's efficacy in accurately interpreting and processing inputs, as well as its overall usability and user experience. This section provides a deep dive into the evaluation methodology employed, shedding light on the robustness and reliability of the system under diverse scenarios and conditions. Through rigorous evaluation, the system's strengths and limitations were elucidated, paving the way for informed conclusions and recommendations for future enhancements and iterations.

After researching the problem and successfully creating a working prototype as a browser plugin, along with getting good evaluation results, the system can keep improving. Author can fix any problems find and add more features to make it better. This way, the system will keep getting better and more useful over time.

5.2 Risks & Limitation

This approach based on the pre-training model which is the primary input in order to predict the required result of the system. Creating a huge collection of word which maps to relevant classes is time consuming. Not only the English word but also training that English word require considerable amount of sentences to predict the meaning correctly. That is major problem in the implementation of the system because currently there is no publicly available training model to detect Sinhala meaning when providing an English word.

System caters without making the approach more complex by not considering the three facts described on the methodology section. First fact is removing the unnecessary meanings for that word. Example, the word “mine” Sinhala/English lexicon which was the basic building block of the system gives “ආකරය, පතල, පතලය, පතල් බහිනවා etc..” as Sinhala meanings. Therefore all of the definitions are not considered. The second one is consider only one selected definition. For above word “mine” the proposed system the author only consider “පතල” as the definition by making it simpler and more user-friendly for readers to view. The last fact indicates that it is focusing only some of the areas such as nouns, pro-nouns and verbs. Otherwise based on the tense of the sentence, the gender wise the definitions differ.

The pre-trained model is assuming the max length of the sentence or the maximum tokens can be found in the sentence is forty. This can be increased according to the requirement. And the limited vocabulary were includes in the pre-trained model.

A properly functioning system enabling a browser plugin will help Sri Lankans to easily understand the meaning of the English word and improve their vocabulary unintentionally while reading the English based content on web pages or web sites they refer.

5.3 Lessons Learned

5.3.1. Sinhala/English language and its importance

Language plays a major role in communication. Through language a person can communicate with the world, define identity, learn, and express history and culture and many more. Learning native language is the building block for expressing everything a person need but as that person grow they communicate with the society, acquire knowledge from various sources. Therefore when considering about the learning a language a person put some extra effort to achieve the concept of learning. When acquiring knowledge the sources not always present from native language. As a Sri Lankan most of the people know Sinhala language but in order to communicate with people or other relevant sources like web sites, office work or academic studies it is very import to have good knowledge of English language because it is a universal language. Therefore people who are acquiring knowledge from web sites or web pages should have a good level of English knowledge. Any language is having ambiguity and this can solve based on the content. Therefore this system would be good solution to get the hundred percent benefits from the web page or web site for the readers to understand the meaning of the word while unintentionally improving their vocabulary.

5.3.2. Use Libraries having latest technologies

Found powerful efficiencies through the adoption of popular open source development libraries for python language such as BERT. For the implementation purpose," Google Colab" was used and this is a well-known product of Google Company and this is heavily used for python application development because user do not need to have special resources and it is having free tier with GPU. With an adequately large user base, online support and documentation is extended through active user forums and published articles.

5.3.3. Practical Difficulties

From the starting in the project many difficulties raised up. Collecting the domain knowledge, proper system design and implementation had many challenges and difficulties. In the evaluation process in the real environment was very difficult. Making pre-trained model was bit hard.

5.3.4 Target audience and Importance of usability design

Usability of a system is a very important factor that should be taken in to account when a product is developed. If the user finds it as difficult then they will reject the product although the system contains very important features. Simply, the ideal solution may not be the best solution for a specific problem. It is very important to identify the target audience and set of clear requirements in the very beginning of the project. Otherwise the total system will be useless, unless these factors are not considered.

5.3.5. Knowledge improvement on best practices

While working on the project, a good understanding on best practices and importance of them were clearly understood. This helped to improve both cording standards and documentation skills. By writing different documents based on the guide lines improved the skills and qualities of document writing. Most importantly following the process helped for a successful completion of implementation of project and thesis. This takes considerable amount of time.

5.4 Future Work

5.4.1. Increase the word collection in the training model

The goal of the project is to identify the ambiguity among English language and provide a solution for making it clear for the readers to distinguish the different meaning based on the context and it is very important. For proving better and accurate predictions it is always good to have a huge collection of words for the training model. Crowdsourcing technique is a good approach for this system to improve for the rest of the English words.

5.4.2. Extend the focus are of words

As mentioned in earlier chapters the proposed system is cater to predict meanings for noun, pronouns and verbs. But it can be extended by considering the tense of the sentence and gender. This is part is important for the completion of the proposed browser plugin.

5.4.3. Multiple Language support

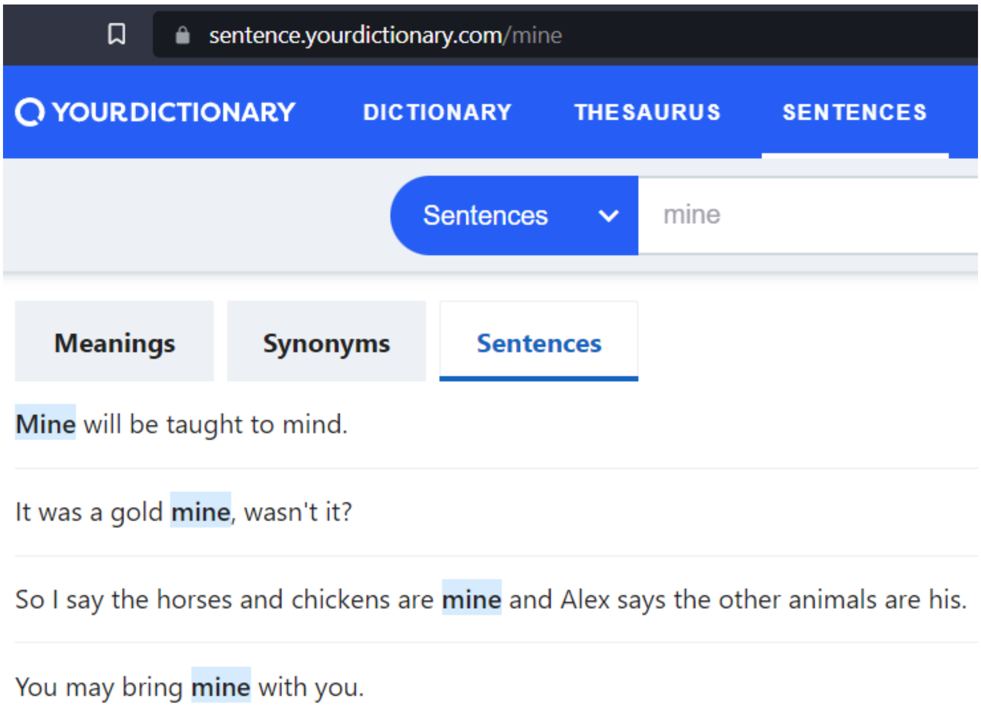
System can be developed for proving meanings according to the language they prefer. There are different languages uses in Sri Lanka. Apart from Sinhala and English Tamil language is another main language which use in Sri Lanka. Using the same approach can cater in Tamil language and that will be a good for Tamil community to improve the English language when they acquiring knowledge from online sources. This also opens for future development.

With these implementations it is possible to have proper functioning system for Sri Lankans to improve English language while referring the online sources.

APPENDICES

To create training model, “Sentence.yourdictionary” was used where the user can enter required word and it display sentences from their sentence repository.

Figure 5.1: Generated mine sentences



Mine will be taught to mind.

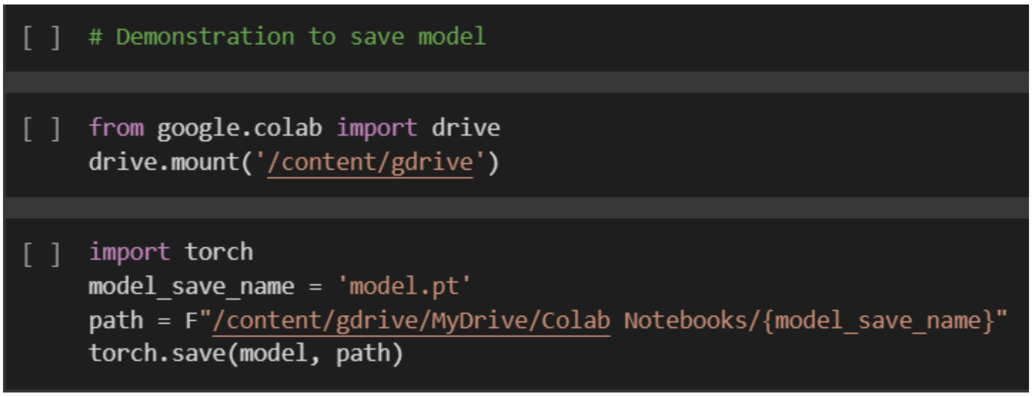
It was a gold mine, wasn't it?

So I say the horses and chickens are mine and Alex says the other animals are his.

You may bring mine with you.

The implemented system uses a training model to predict the correct Sinhala definition for the selected English word in the sentence. Machine requires lot of resources for training these types of training models. Pre-trained models can use for resource utilization and reducing the time taken to train the model. Below figure illustrates the code segment for saving the model inside Google drive.

Figure 5.2: Code segment to save trained model



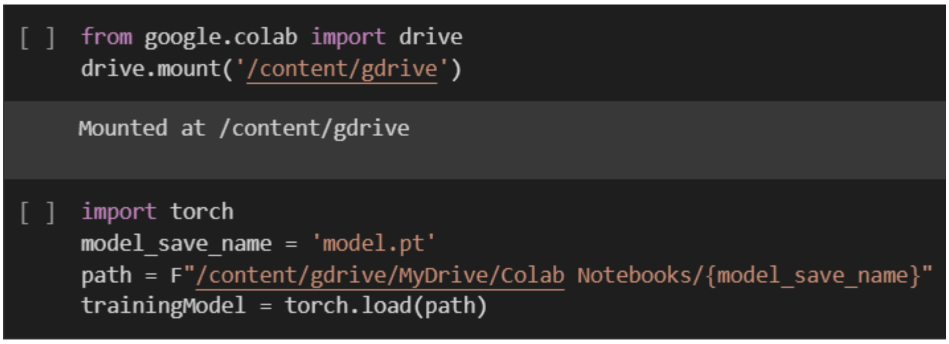
```
[ ]  # Demonstration to save model

[ ]  from google.colab import drive
     drive.mount('/content/gdrive')

[ ]  import torch
     model_save_name = 'model.pt'
     path = F"/content/gdrive/MyDrive/Colab Notebooks/{model_save_name}"
     torch.save(model, path)
```

Pre trained model can then use in the implementations. Below figure illustrates the code segment required for that purpose.

Figure 5.3: Code segment to retrieve pre-trained model



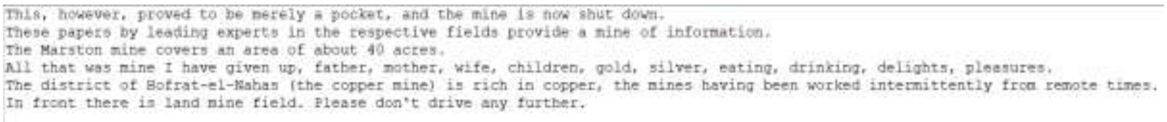
```
[ ] from google.colab import drive
    drive.mount('/content/gdrive')

    Mounted at /content/gdrive

[ ] import torch
    model_save_name = 'model.pt'
    path = F"/content/gdrive/MyDrive/Colab Notebooks/{model_save_name}"
    trainingModel = torch.load(path)
```

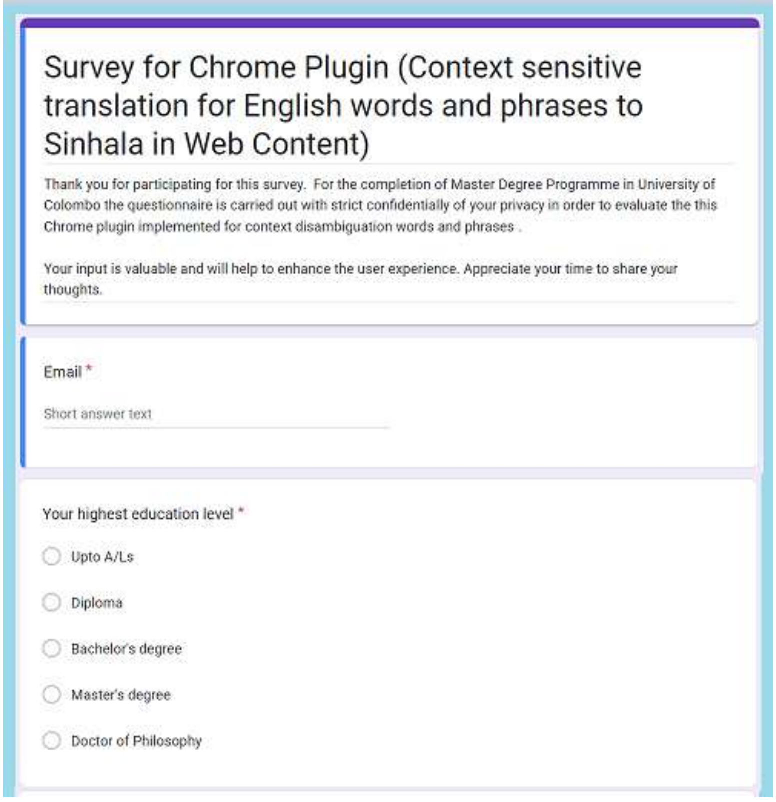
Following figure illustrates the sentences which used for the evaluation process.

Figure 5.4: Sentences used for evaluation process



```
This, however, proved to be merely a pocket, and the mine is now shut down.
These papers by leading experts in the respective fields provide a mine of information.
The Marston mine covers an area of about 40 acres.
All that was mine I have given up, father, mother, wife, children, gold, silver, eating, drinking, delights, pleasures.
The district of Bofrat-el-Nahas (the copper mine) is rich in copper, the mines having been worked intermittently from remote times.
In front there is land mine field. Please don't drive any further.
```

Figure 5.5: Survey form



Survey for Chrome Plugin (Context sensitive translation for English words and phrases to Sinhala in Web Content)

Thank you for participating for this survey. For the completion of Master Degree Programme in University of Colombo the questionnaire is carried out with strict confidentially of your privacy in order to evaluate the this Chrome plugin implemented for context disambiguation words and phrases .

Your input is valuable and will help to enhance the user experience. Appreciate your time to share your thoughts.

Email *

Short answer text

Your highest education level *

○ Upto A/Ls

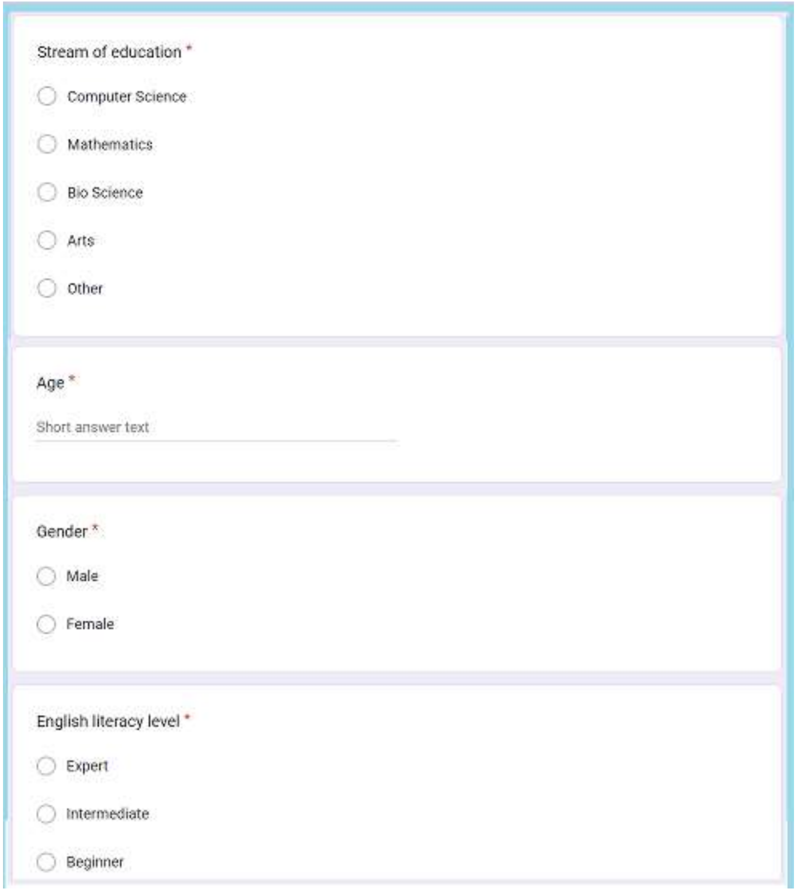
○ Diploma

○ Bachelor's degree

○ Master's degree

○ Doctor of Philosophy

Figure 5.6: Survey form



Stream of education *

○ Computer Science

○ Mathematics

○ Bio Science

○ Arts

○ Other

Age *

Short answer text

Gender *

○ Male

○ Female

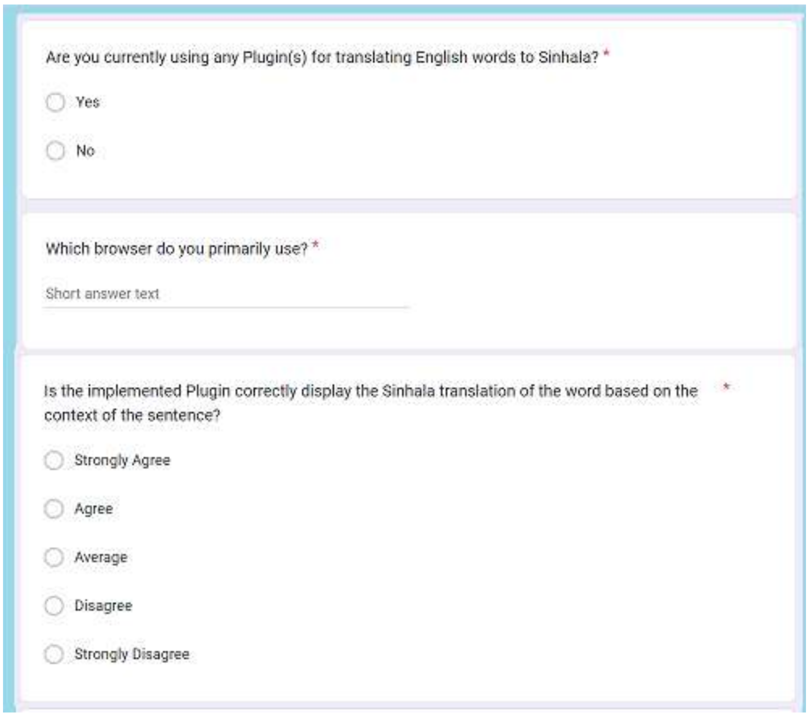
English literacy level *

○ Expert

○ Intermediate

○ Beginner

Figure 5.7: Survey form



Are you currently using any Plugin(s) for translating English words to Sinhala? *

○ Yes

○ No

Which browser do you primarily use? *

Short answer text

Is the implemented Plugin correctly display the Sinhala translation of the word based on the context of the sentence? *

○ Strongly Agree

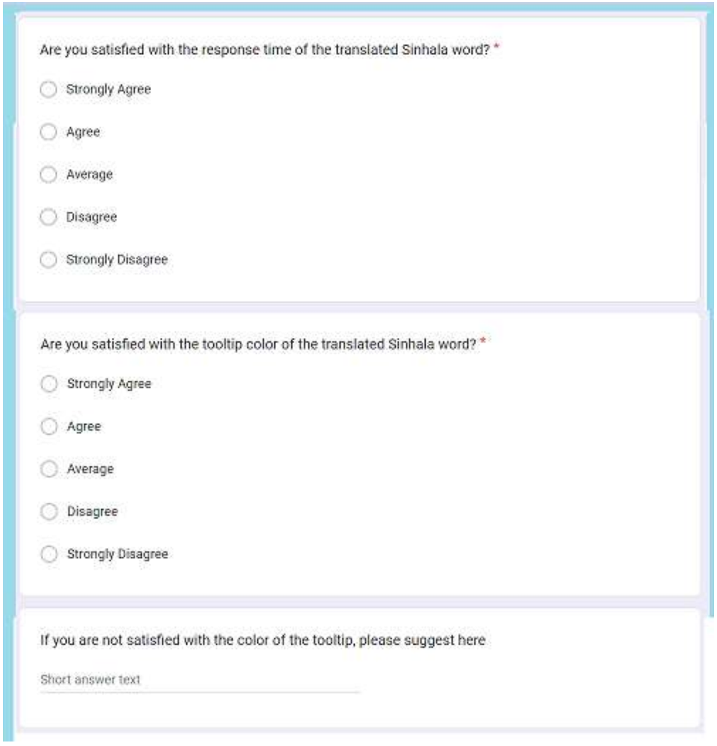
○ Agree

○ Average

○ Disagree

○ Strongly Disagree

Figure 5.8: Survey form



Are you satisfied with the response time of the translated Sinhala word? *

○ Strongly Agree

○ Agree

○ Average

○ Disagree

○ Strongly Disagree

Are you satisfied with the tooltip color of the translated Sinhala word? *

○ Strongly Agree

○ Agree

○ Average

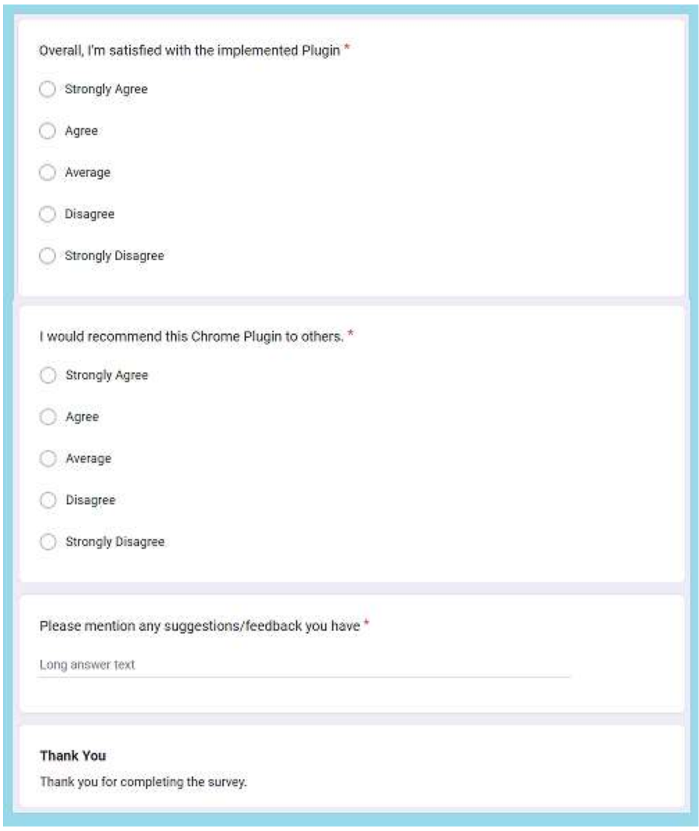
○ Disagree

○ Strongly Disagree

If you are not satisfied with the color of the tooltip, please suggest here

Short answer text

Figure 5.9: Survey form



Overall, I'm satisfied with the implemented Plugin *

○ Strongly Agree

○ Agree

○ Average

○ Disagree

○ Strongly Disagree

I would recommend this Chrome Plugin to others. *

○ Strongly Agree

○ Agree

○ Average

○ Disagree

○ Strongly Disagree

Please mention any suggestions/feedback you have *

Long answer text

Thank You

Thank you for completing the survey.

Table 5.1 : Random 100 sentences checked for the evaluation and results

	Sentence	Word	Predicted Sinhala translation	Correct (Y / N)
1	She deposited her paycheck at the bank.	bank	බැංකුව	Y
2	The canoe bumped against the river bank.	bank	ඉවුර | බැංකුව	N
3	The dog's bark echoed through the forest.	bark	බුරනවා	Y
4	The tree's bark was rough to the touch.	bark	පොත්ත	Y
5	He programmed in Python for his project.	python	ක්රමලේඛන භාෂාව	Y
6	The python slithered through the grass.	python	පිඹුරා	Y
7	The bear emerged from its den in the woods.	bear	වලහා	Y
8	She couldn't bear the thought of leaving.	bear	දරාගන්නවා	Y
9	The seal basked in the sun on the rocky shore.	seal	සීල් මත්ස්යයා	Y
10	He used a seal to close the envelope.	seal	මුද්රාව	Y
11	A crane soared gracefully above the marsh.	crane	කොකා	Y
12	The construction crew used a crane to lift heavy materials.	crane	දොඹකරය	Y
13	The bat flew silently through the night.	bat	වවුලා	Y
14	He swung the cricket bat with precision.	bat	පිත්ත	Y
15	The soccer match ended in a tie.	match	තරඟය	Y
16	She found her perfect match in him.	match	තරඟය	N
17	She wore a beautiful ring on her finger.	ring	මුද්ද	Y
18	The sound of the ring echoed in the hallway.	ring	මුද්ද	N
19	She bowed after her performance.	bow	ආචාර කරනවා	Y
20	He aimed his bow at the target.	bow	දුන්න	Y
21	The mine was rich in precious metals.	mine	පතල	Y
22	She claimed the discovery of a gold mine.	mine	පතල	Y
23	He washed his face with soap and water.	face	මුහුණ	Y
26	She had a stern face during the meeting.	face	මුහුණ	Y
25	The duck paddled gracefully across the pond.	duck	තාරාවා	Y
26	She wore a dress made of duck fabric.	duck	දළ රෙදි වර්ගයක්	Y
27	He gave a quick duck of his head in agreement.	duck	හිස නැමීම	Y
28	He faced his fears and conquered them.	face	මුහුණ	Y
29	She couldn't face the truth.	face	මුහුණ පානවා	Y
30	The duck quacked loudly at the pond.	duck	තාරාවා	Y
31	She wore a duck brooch on her jacket.	duck	තාරාවා	Y
32	He ducked under the low-hanging branch.	duck	හිස නැමීම	Y
33	She faced her opponent in the final match.	face	මුහුණ | මුහුණ පානවා	N
34	The face of the clock showed the time.	face	මුහුණ	Y
35	The child's face lit up with excitement.	face	මුහුණ	Y
36	The bear lumbered through the forest.	bear	වලහා	Y
37	She couldn't bear to see him leave.	bear	දරාගන්නවා	Y
38	The seal dove into the ocean.	seal	සීල් මත්ස්යයා	Y
39	He used a seal to mark the letter as confidential.	seal	මුද්රාව	Y
40	The crane stood tall in the wetlands.	crane	කොකා	Y
41	They used a crane to hoist the cargo onto the	crane	දොඹකරය	Y

	ship.			
42	The bat hung upside down from the branch.	bat	වවුලා	Y
43	He swung the bat and hit a home run.	bat	පිත්ත	Y
44	The soccer match drew a large crowd.	match	තරඟය	Y
45	She found a perfect match for her shoes.	match	ගැළපීම	Y
46	She wore a diamond ring on her finger.	ring	මුද්ද	Y
47	The church bells began to ring.	ring	නාද කරනවා	Y
48	She gave a slight bow to the audience.	bow	ආචාර කරනවා	Y
49	He aimed his bow at the target.	bow	දුන්න	Y
50	The mine collapsed, trapping the workers inside.	mine	පතල	Y
51	She explored the depths of the gold mine.	mine	පතල	Y
52	He washed his face with cold water.	face	මුහුණ	Y
53	She had a friendly face that made everyone feel welcome.	face	මුහුණ	Y
54	The ducklings followed their mother across the pond.	duck	Not found	N
55	She sewed a dress from duck fabric.	duck	දළ රෙදි වර්ගයක්	Y
56	He gave a quick duck to avoid hitting his head.	duck	හිස නැමීම	Y
57	He faced his fears and overcame them.	face	මුහුණ පානවා	Y
58	She couldn't face the consequences of her actions.	face	මුහුණ පානවා	Y
59	The duck quacked loudly, demanding attention.	duck	තාරාවා	Y
60	She admired the sunset from the elevated bank overlooking the valley.	bank	ඉවුර | බැංකුව	N
61	He ducked behind the tree to hide from the rain.	duck	හිස නැමීම	Y
62	She faced her opponent in the final round.	face	මුහුණ | මුහුණ පානවා	N
63	The clock face was adorned with Roman numerals.	face	මුහුණ	Y
64	Her face lit up with joy when she saw him.	face	මුහුණ	N
65	The bear roamed the forest in search of food.	bear	වලහා	Y
66	She couldn't bear the thought of losing him.	bear	දරාගන්නවා	Y
67	The seal lounged on the rocky beach.	seal	සීල් මත්ස්‍යයා	Y
68	He used wax to seal the envelope shut.	seal	මුද්‍රාව	Y
69	The crane soared gracefully through the air.	crane	කොකා	Y
70	They used a crane to lift the heavy machinery.	crane	දොඹකරය	Y
71	The bat flew silently in the night sky.	bat	වවුලා	Y
72	He swung the cricket bat with all his might.	bat	පිත්ත	Y
73	The soccer match ended in a tie score.	match	තරඟය	Y
74	She found her perfect match in him.	match	ගැළපීම	Y
75	She wore a ring on her finger.	ring	මුද්ද	Y
76	With each tap of the chisel against the stone, a clear ringing sound reverberated in the sculptor's studio.	ring	මුද්ද	N
77	She gave a slight bow to the audience.	bow	ආචාර කරනවා	Y
78	He aimed his bow at the target.	bow	දුන්න	Y
79	The mine was closed due to safety concerns.	mine	පතල	Y

80	She discovered a rich vein of gold in the mine.	mine	පතල	Y
81	He washed his face with warm water.	face	මුහුණ	Y
82	She had a kind face that put people at ease.	face	මුහුණ	Y
83	The duck paddled gracefully across the lake.	duck	තාරාවා	Y
84	She wore a coat made of duck fabric.	duck	දළ රෙදි වර්ගයක්	Y
85	He gave a quick duck to avoid the low branch.	duck	හිස නැමීම	Y
86	The river bank was dotted with wildflowers in bloom.	bank	ඉවුර | බැංකුව	N
87	She couldn't face the truth about what happened.	face	මුහුණ පානවා	Y
88	The duck quacked loudly, demanding attention.	duck	තාරාවා	Y
89	She wore a duck shaped pendant around her neck.	duck	දළ රෙදි වර්ගයක්	N
90	He ducked under the door frame to enter the room.	duck	හිස නැමීම	Y
91	She faced her opponent in the final round.	face	මුහුණ පානවා	Y
92	The clock face was adorned with intricate designs.	face	මුහුණ	Y
93	Her face lit up with joy when she saw him.	face	මුහුණ	Y
94	The bear wandered through the forest in search of food.	bear	වලහා	Y
95	She couldn't bear the thought of losing him.	bear	දරාගන්නවා	Y
96	The seal basked in the warm sun on the rocks.	seal	සීල් මත්ස්‍යයා	Y
97	He used tape to seal the box shut.	seal	මුද්‍රාව	Y
98	The crane stood tall against the skyline.	crane	දොඹකරය	Y
99	They used a crane to lift the heavy load.	crane	දොඹකරය	Y
100	The bat flew silently through the night sky.	bat	වවුලා	Y

REFERENCES

A Comprehensive Study on Sinhala and English Verbs, 2018. . Int. J. Stud. Engl. Lang. Lit. 6. https://doi.org/10.20431/2347-3134.0609006.

Amano, S., 1987. TAURAS: The Toshiba Machine Translation System. undefined.

Babelfish.com [WWW Document], n.d. URL https://www.babelfish.com/ (accessed 11.14.22).

BERT Word Embeddings Tutorial · Chris McCormick [WWW Document], n.d. URL https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/ (accessed 11.15.22).

Bharati, A., Chaitanya, V., Sangal, R., Gillon, B., 2002. Natural Language Processing: A Paninian Perspective.

Chaudhury, S., Rao, A., Sharma, D.M., 2010. Anusaaraka: An expert system based machine translation system, in: Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010). Presented at the 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), IEEE, Beijing, China, pp. 1–6. https://doi.org/10.1109/NLPKE.2010.5587789

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Presented at the NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Hettige, B., Karunananda, A., Rzevski, G., n.d. Selected Text Machine Translator for English to Sinhala 6.

Hettige, B., Karunananda, A.S., 2010. Varanageema: A Theoretical basics for English to Sinhala Machine Translation 8.

Hettige, B., Karunananda, A.S., 2009. On Demand Web Page Translation -BEES in Action-9.

Hettige, B., A. S. Karunananda, and G. Rzevski. 2016. "A Multi-Agent Solution for Managing Complexity in English to Sinhala Machine Translation." *International Journal of*

Design & Nature and Ecodynamics 11 (2): 88–96. https://doi.org/10.2495/DNE-V11-N2-88-96.

Horev, R., 2018. BERT Explained: State of the art language model for NLP [WWW Document]. Medium. URL https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (accessed 11.15.22).

Jayasuriya, M., Weerasinghe, R., 2013. Learning a stochastic part of speech tagger for Sinhala. Presented at the International Conference on Advances in ICT for Emerging Regions, ICTer 2013 - Conference Proceedings, pp. 137–143. https://doi.org/10.1109/ICTer.2013.6761168

Lavie, A., Vogel, S., Levin, L., Peterson, E., Probst, K., Llitjós, A.F., Reynolds, R., Carbonell, J., Cohen, R., 2003a. Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. ACM Trans. Asian Lang. Inf. Process. 2, 143–163. https://doi.org/10.1145/974740.974747

Lavie, A., Vogel, S., Levin, L., Peterson, E., Probst, K., Llitjós, A.F., Reynolds, R., Carbonell, J., Cohen, R., 2003b. Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. ACM Trans. Asian Lang. Inf. Process. 2, 143–163. https://doi.org/10.1145/974740.974747

Mandal, D., Gupta, M., Dandapat, S., Banerjee, P., Sarkar, S., 2007. Bengali and Hindi to english CLIR evaluation. pp. 95–102. https://doi.org/10.1007/978-3-540-85760-0_12

Medium [WWW Document], n.d. . Medium. URL https://towardsdatascience.com/identifying-the-right-meaning-of-the-words-using-bert-817eef2ac1f0. (accessed 11.15.22).

Németh, G.D., 2019. Identifying the right meaning of the words using BERT [WWW Document]. Medium. URL https://towardsdatascience.com/identifying-the-right-meaning-of-the-words-using-bert-817eef2ac1f0 (accessed 11.15.22).

Sentence Examples | Examples of Words Used in a Sentence [WWW Document], n.d. URL https://sentence.yourdictionary.com/ (accessed 11.15.22).

Shalini, R.M.M., Hettige, B., 2017. Dictionary Based Machine Translation System for Pali to Sinhala. Sri Lanka 6.

Sjöberg, A., n.d. The Use Of The Copula In Non-Copula Constructions In The Languages Of South Asia 89.

Soudi, A., Cavalli-Sforza, V., Jamari, A., n.d. A Prototype English-to-Arabic Interlingua-based MT system 5.

Sripirakas, S., Weerasinghe, A.R., Herath, D.L., 2010. Statistical machine translation of systems for Sinhala - Tamil, in: 2010 International Conference on Advances in ICT for Emerging Regions (ICTer). Presented at the 2010 International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE, Colombo, Sri Lanka, pp. 62–68. https://doi.org/10.1109/ICTER.2010.5643268

Understanding LSTM Networks -- colah's blog [WWW Document], n.d. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed 11.15.22).

Usage Statistics and Market Share of Content Languages for Websites, November 2022 [WWW Document], n.d. URL https://w3techs.com/technologies/overview/content_language (accessed 11.14.22).

Wasala, A., Weerasinghe, R., n.d. EnSiTip: A Tool to Unlock the English Web 7.

Weerasinghe, A.R., Herath, D.L., Medagoda, N.P.K., 2006. A KNN based Algorithm for Printed Sinhala Character Recognition, in: Proceedings of 8th International Information Technology Conference.

Weerasinghe, R., and Premachandra, A. 2008. "Example Based Machine Translation for English-Sinhala Translations.".

Weerasinghe, R., Herath, D., Welgama, V., 2009. Corpus-based Sinhala lexicon 17–23. https://doi.org/10.3115/1690299.1690302

Weerasinghe, R., Wasala, A., Gamage, K., 2005. A Rule Based Syllabification Algorithm for Sinhala. pp. 438–449. https://doi.org/10.1007/11562214_39

Yates, and Sarah. 2006. "Babel Fish (An Analysis of the Machine Translation of Legal Information)," 3, 98 (June):481–500