Analysis of Sinhala Using Natural Language Processing Techniques

Sajika Gallege

Department of Computer Sciences
University of Wisconsin-Madison
1210 W. Dayton Street, Madison, WI 53706
sgallege@cs.wisc.edu

Abstract

Sinhala is the native language of the island nation of Sri Lanka. It belongs to the Indo-Aryan branch of the Indo-European languages. Sinhala has a written alphabet which consists of 54 basic characters. In my project I have applied some of the Natural Language Processing (NLP) techniques to analyze the Sinhala language to gain a better understanding of the language in a NLP perspective and as a step towards developing more complex tools for machine translation, spelling/ grammar correction and speech recognition. The first step of the project was to collect a sufficient text corpus and to pre-process the text to apply the NLP algorithms. The experiments performed include Maximum Likelihood Estimates (MLE) on Sinhala Characters, Language Identification using a Naïve Bayes Classifier, Zipf's Law Behavior, Topic Classification using Support Vector Machines (SVM) and Language Models. All of the NLP techniques applied to the collected corpus produced satisfactory results. This is an encouraging start for further research on the Sinhala language.

Introduction

The Sinhala Language

Sinhala is the native language of the island nation of Sri Lanka. It belongs to the Indo-Aryan branch of the Indo-European languages. Sinhala is the mother tongue of about 15 million Sinhalese, while it is spoken by about 19 million people in total. The oldest Sinhala inscriptions found are from the third or second centuries BCE; the oldest existing literary works date from the ninth century CE.

The Sinhala Alphabet

Sinhala has a written alphabet which consists of 54 basic characters. Sinhala sentences are written from left to right. Most of the Sinhala letters are curlicues.

The Sinhala alphabet consists of 18 vowel characters and 36 consonant characters. The vowels include 8 stops, 2 fricatives, 2 affricates, 2 nasals, 2 liquids and 2 glides.

The Unicode range for Sinhala is U+0D80–U+0DFF. The code page can be found at www.unicode.org /charts/PDF/U0D80.pdf. Given below is the Unicode mapping of the Sinhala alphabet

	0D8x	0D9x	0DAx	0DBx	0DCx	0DDx	0DEx	0DFx
0		පෟ	ච	ධ	ව	◌ෘ		
1		ඵ	ජ	න	ශ	◌ෲ		
2	◌ං	බ	ඣ		ෂ	◌ෳ		◌aa
3	◌ඃ	භ	ඤ	ඳ	ස	◌ෟ		◌෴
4		ම	ඥ	ප	හ	◌ෟ		◌෵
5	අ	ඹ	ඦ	ඵ	ළ			
6	ආ	ය	ඣ	බ	ෆ	◌ෟ		
7	ඇ		ට	භ				
8	ඈ		ඨ	ම		◌ෙ		
9	ඉ		ඩ	ඹ		◌ෙ◌		
A	ඊ	ක	ඪ	ය	◌්	◌ෙ		
B	උ	ඛ	ණ	ර		◌ෙෙ		
C	ඌ	ග	ඬ			◌ො		
D	ඍa	ස	ත	ල		◌ෝ		
E	ඍaa	ඬ	ථ			◌ෞ		
F	ඎ	හ	ද		◌ා	◌෴		

Related Work

The Language Technology Research Laboratory (LTRL) of The University of Colombo School of Computing has been involved in Sinhala language related NLP research since 2004. The research work conducted by LTRL includes producing a large Sinhala Corpus, a Lexical Resource, a Text-to-Speech Engine (TTS) and an Optical Character Recognition application (OCR).

The Corpus and Pre-processing

The text corpus collected for this project has 681 233 word tokens, 74 369 word types, and 2 268 895 basic Sinhala characters.

The corpus consists of documents from several categories. The main categories are news articles, sports articles, feature articles, short stories, poems, news headlines, and sports headlines. The news, sports and feature documents make up about 70 percent of the corpus, while the other categories make up the balance 30 percent.

The following sources were used to collect text for the corpus: LTRL Sinhala corpus www.ucsc.cmb.ac.lk/ltrl/, stories by Martin Wickramasinghe www.martinwickramasinghe.org, and online newspapers www.divaina.com, www.silumina.lk, www.lankadeepa.lk, www.defence.lk/sinhala.

Collecting a sufficient text corpus was an important part of the project and it was challenging due to several reasons. First of all, the Sinhala text content available over the internet is limited, and the available content is not consistent because different web sites use different text encodings and fonts. This challenge was overcome by collecting articles from newspaper website archives and using the Unicode character encoding tool from the LTRL. The second challenge was that many of the NLP tools only support ASCII encoding, but Sinhala text uses Unicode. This was overcome by pre processing the text to suit each of the algorithms. Specific pre processing steps for each test is given under the tests. In pre processing most of the non Sinhala characters were removed for simplicity.

The NLP Analysis of Sinhala

1. Maximum Likelihood Estimate (MLE) on Sinhala Characters

The goal of the test was to observe the MLE's of the characters in the collected corpus and to observe which characters are most frequent in Sinhala.

Dataset: The whole text corpus was used for calculating MLE's.

Pre processing: For simplicity, only the counts of main Sinhala characters were considered. All non Sinhala characters and punctuation were ignored. Two versions of the test were run with and without the inclusion of the white space.

Algorithm: Maximum Likelihood Estimate is defined as

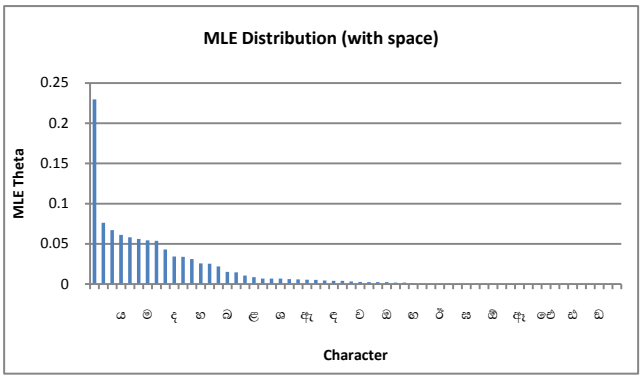
$$MLE\ \theta = \frac{n_c}{N}$$

Where n_c is the count of a particular character and N is the total number of characters in the corpus. To obtain the counts, the Corpus is traversed once while maintaining a counter for each character.

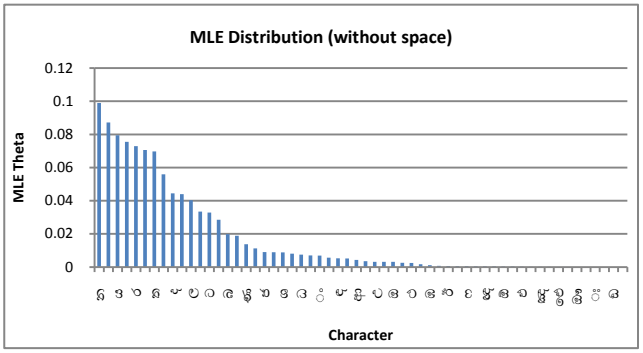
Results: The ten most frequent characters are listed together with the counts and MLE estimate in the table below.

Char	Count	MLE
	676085	0.229572017
ත	224464	0.076219193
ව	197772	0.067155634
ය	180277	0.061215017
ක	171259	0.058152857
ර	165380	0.056156578
ම	160238	0.054410556
න	158262	0.053739584
ස	127016	0.043129665
ද	100910	0.034265088

The following chart displays the distribution of the MLE for the characters with the white space included.



The following chart displays the distribution of the MLE for the characters without the white space.



Conclusion: White space seems to be the most frequent character in the corpus and it seems to appear about three times more frequently than the next character 'න' in the list. It is also noteworthy that none of the vowels are among the top ten (the first vowel 'ෘ' is at the 16th position). This could be because in Sinhala the vowel sounds are added as an add-on modifier to a consonant, instead of as a new character. In this experiment we only counted the basic characters, disregarding any add-ons.

2. Language Identification Using a Naïve Bayes Classifier

The goal of the test was to check the effectiveness of Naïve Bayes language identifier in classifying Sinhala against English, Spanish, and Japanese.

Dataset: The Sinhala dataset consists of 20 feature articles from online newspapers (www.silumina.lk). The English, Spanish and Japanese documents were obtained from http://pages.cs.wisc.edu/jerryzhu/cs769/dataset/languageID.tgz.

Pre processing: The Sinhala text was converted to English text, by replacing each character with a corresponding English syllable. Sinhala phrases written using English characters are informally known as 'Singlish'

eg: දිලිසෙන සියල්ල රත්තරන් නොවේ ➔
dhilisena siyalla raththaran novea

Algorithm: To find the most likely language given a document we need to calculate the maximum conditional probability defined as

$$P(Language|Document) = P(Document \mid Language) \,.\, P(Language)$$

The prior probabilities are calculated using:

$$P(Language) = \frac{Number\ of\ Documents\ in\ Language}{Total\ Documents}$$

By the Naïve Bayes assumption we have:

$$P(Document \mid Language) \approx \prod_{i=1}^{n} P(c_i|Languge)$$

Conditional Likelihoods are calculated as:

$$P(ci|\ Language) = \frac{countLanguage(ci)}{nCharLanguage}$$

Where countLanguage(c_i) is the number of times character c_i occurs in all particular language documents in the training set.

All probabilities were converted to log to avoid underflow and add 1 smoothing was used.

Sinhala Conditional Probabilities:

P(a|Sinhala) = 0.26629795758393937
P(b|Sinhala) = 0.01064756347167182
P(c|Sinhala) = 9.362888124373993E-4
P(d|Sinhala) = 0.02939511387884858
P(e|Sinhala) = 0.04576928101728868
P(f|Sinhala) = 2.6128990114532074E-4
P(g|Sinhala) = 0.013434655750555241
P(h|Sinhala) = 0.07483778251970562
P(i|Sinhala) = 0.06675956974262945
P(j|Sinhala) = 0.004572573270043113
P(k|Sinhala) = 0.031899142098157904
P(l|Sinhala) = 0.018072551495884683
P(m|Sinhala) = 0.031289465662152155
P(n|Sinhala) = 0.055001524191090015
P(o|Sinhala) = 0.010233854461525062
P(p|Sinhala) = 0.016679005356442973
P(q|Sinhala) = 2.177415842877673E-5
P(r|Sinhala) = 0.03033140269128598
P(s|Sinhala) = 0.031899142098157904
P(t|Sinhala) = 0.04378783260027
P(u|Sinhala) = 0.03081043417671907
P(v|Sinhala) = 0.03710316596263554
P(w|Sinhala) = 1.9596742585899056E-4
P(x|Sinhala) = 2.177415842877673E-5
P(y|Sinhala) = 0.031049949919435615
P(z|Sinhala) = 2.177415842877673E-5
P(|Sinhala) = 0.11866916343683316

A test document classified as Sinhala if
log P(Sinhala | doc) > log P(English | doc) and
log P(Sinhala | doc) > log P(Spanish| doc) and
log P(Sinhala | doc) > log P(Japanese| doc).
The same procedure is followed for other languages

Results: In the form of a confusion matrix

	True Sinhala	True English	True Spanish	True Japanese
Predicted as Sinhala	10	0	0	0
Predicted as English	0	10	0	0
Predicted as Spanish	0	0	10	0
Predicted as Japanese	0	0	0	10

Conclusion: It is evident from the confusion matrix that all the documents are classified correctly without any false positives or false negatives. The Naïve Bayes language classifier accurately classifies Sinhala apart from English, Spanish and Japanese with 100 percent accuracy.

3. Zipf's Law Behavior

The goal of this test was to observe if Sinhala displays the Zipf's Law behavior. Zipf's Law states that, given a text corpus, if *f*: is word count and *r*: is rank, when sorted by word count that

$$f.r \approx Constant$$

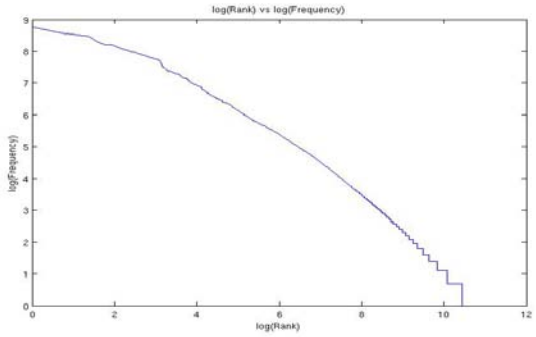
Dataset: The whole text corpus was used for calculating word counts.

Pre processing/ Algorithm: The whole text corpus was merged into a single document. Then, the document was traversed while counting how many times each word appears. Finally, the list was sorted by the count in the descending order and the rank was assigned.

Results: The top ten words of the sorted list are given below. The English translations of the words are also listed. Please note that some of the meanings of some Sinhala words change depending on the context, so the given translation may not be exact.

Word	Translation	*f*	*r*
ද	and/also	6467	1
මේ	this	5321	2
ය	the	5015	3
හා	and/with	4805	4
ඒ	that	3954	5
ම	a	3684	6
ඇති	has	3663	7
බව	about	3346	8
දී	at	3166	9
වන	is/of	3064	10

Given below is a plot of *log(r)* versus *log(f)*



Conclusion: From the above graph we can observe that the words roughly form a line from the upper-left corner to the lower-right corner of the graph. This indicates that the Sinhala corpus displays Zipf's Law behavior. Looking at the sorted list of words we can conclude that the top ranked words are stop words. This shows that developing a stop word removal algorithm for Sinhala might be beneficial for NLP purposes.

4. Topic Classification Using Support Vector Machines (SVM)

The goal of this experiment was to test the effectiveness of SVM in Sinhala topic classification. Two sets of topics are used in this experiment. The first classification was on sports versus news, and the second classification was on 2009 news versus 2010 news. Both linear and polynomial SVM kernels were used for the classification tasks to determine which kernel performs better.

Dataset: The dataset consists of four parts, two for each classification task. For the News versus Sports classification, there are 500 news headlines and 500 sports headlines. The data was collected from http://www.divaina.com/ archive on randomly picked dates from 2009 and 2010.

For the 2009 News versus 2010 News classification there are 500 news headlines from 2009 and 500 news headlines from 2010. The data was collected from http://www.divaina.com/ archive on randomly picked dates between January and June from years 2009 and 2010. This is an interesting comparison because of the major events that took place in Sri Lanka in 2009 and 2010. The year 2009 saw an end to a 30 year old terrorist insurgency, so the news from 2009 is expected to have more defense related headlines. In 2010 a presidential election and a general election took place, so the news from 2010 is expected to have more political content.

Pre processing: The first step was to combine all the headlines from a classification task to create a vocabulary. Then each headline was converted into a Bag of Words (BOW) vector with the class label (+1/-1)

eg: සූකර උණ තවත් බිල්ලක් ගනියි →

-1 116:1.0 211:1.0 212:1.0 3622:1.0 4548:1.0

Next the BOW vectors from +/- classes were randomly picked to create 10 train/ test folds, such that the test set consists of 10 percent of the data (100 headlines) and the train set consists of 90 percent of the data (900 headlines).

Algorithm: The SVM creates a hyper plane in the middle of the two classes, so that the distance to the nearest positive or negative example is maximized.

$$max_{w,b} \frac{1}{||w||} \qquad s.t \quad y_i(w^t x_i + b) \geq 1 \;\; i = 1..n$$

The SVM light software from http://svmlight.joachims.org/ was used for this test. The default linear kernel and polynomial kernel with settings (-s 1 –r 1 –d 1) was used for all the folds.

Results: The first table shows the comparison of test set accuracies from the News versus Sports classification together with the mean, standard deviation and the t-value from the two-tailed paired t-test.

News Vs. Sports		
Fold #	Linear Kernel	Polynomial Kernel
1	94	92
2	87	87
3	90	89
4	94	94
5	92	92
6	89	90
7	86	87
8	90	90
9	88	88
10	91	90
mean	90.1	89.9
st. dev	2.726414	2.282786
t-Value		0.508646

The next table shows the same information for the 2009 News versus 2010 News classification.

2009 News Vs 2010 News		
Fold #	Linear Kernel	Polynomial Kernel
1	88.89	88.89
2	88.89	88.89
3	89.9	87.88
4	91.92	91.92
5	91.92	91.92
6	90.91	90.91
7	87.88	86.87
8	90.91	88.89
9	89.9	86.87
10	93.94	93.94
mean	90.506	89.698
st. dev	1.7941522	2.3710513
t-Value		0.052839

Conclusion: From above results it is evident that the SVM performs well on Sinhala topic classification. The News versus Sports was best classified by the linear kernel with a mean average of 90.1 percent. The 2009 News versus 2010 News was best classified by the linear SVM kernel with 90.5 percent accuracy. The linear SVM kernel performed better on both classification tasks, but the difference between the linear kernel and the polynomial kernel is not statistically significant in either case.

5. Language Model and Perplexity

The goal of this experiment was to generate n-gram Language Models for Sinhala where n=1, 2, 3, 4 and compare the perplexity on the train and test sets. A good Language Model is essential for many advanced NLP tools such as speech recognition and grammar correction.

Dataset: The whole text corpus was used for generating the Language Model

Pre processing: The language modeling tool did not accept Unicode characters so the Sinhala text needed to be converted to a format that would be accepted by the language modeling tool.

The first step was to create a vocabulary from the complete text corpus. Then a unique index was assigned to every word in the vocabulary. Afterwards, each word in the corpus was replaced with the corresponding index.

eg: තවත් ලෝක වාර්තාවක් තැබීම සතුටක්, මුරලි ➔
659 61 1101 1641 1642 319

After the conversion 10 percent of the corpus was set aside as the test set.

Algorithm: An n-gram Language Model is defined as

$$P(w_i|w_{1:i-1}) \approx P(w_i|w_{i-n+1:i-1})$$

The conditioning part $w_{i-n+1:i-1}$ is called 'history', which has n−1 previous words.

Perplexity is defined as

$$PP(C';\theta) = 2^{-\frac{1}{|C'|}\sum_{w=1}^{V} C'_w log_2 \theta_w} = P(C'|\theta)^{-\frac{1}{|C'|}}$$

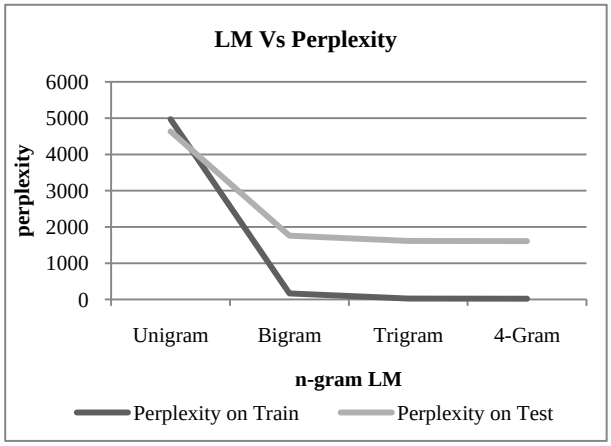
Perplexity measures, on average, how many 'equally likely' words we must choose from for each word position – the smaller the number, the more certain we are, and the better the model θ.

The Statistical Language Modeling Toolkit from http://www.speech.cs.cmu.edu/SLM/toolkit.html was used to generate the Language Models and calculate the perplexities.

Results: The following table shows the train set and test set perplexities from different n-gram Language Models

	Perplexity on Train	Perplexity on Test
Unigram LM	4968.95	4634.47
Bigram LM	166.91	1758.74
Trigram LM	24.87	1614.47
4-Gram LM	20.75	1611.63

A plot showing the perplexities is given below:



Conclusion: The Perplexity seems to reduce as the Language Model gets more complex. There is a drastic reduction of perplexity from unigram to bigram language model. The test perplexity shows a slight reduction for trigram and 4-gram LM's. The big difference between the train set and test set perplexity may be due to overfitting because of the limited corpus size.

Future Work

The NLP analysis on Sinhala provided a good insight to the language. The effectiveness of the tested algorithms encourages further research into the Sinhala language. There are many areas to be researched and many practical applications. Some applications that can be based on NLP

research are machine translation, spelling/ grammar correction and speech recognition.

References

Samaranayaka, V. K., Nandasara, S. T., Dissanayake, J. B., Weerasinghe, A. R., & Wijayawardena, H. (2001). *An Introduction to UNICODE for Sinhala Characters.* Colombo: University of Colombo.

Weerasinghe, R., Hearath, D., & Welgama, V. (2009). *Corpus-based Sinhala Lexicon.* Colombo: University of Colombo.

Wijayawardhana, H. (2001). *Rendering of Unicode Sinhala Characters.* Colombo: University of Colombo.

Zhu, X. (2010). *CS769 Spring 2010 Advanced Natural Language Processing.* Madison: University of Wisconsin-Madison.