Single-Document Myanmar Text Summarization using Latent Semantic Analysis (LSA)

Soe Soe Lwin*, Dr. Khin Thandar Nwet**
*University of Computer Studies, Yangon(UCSY), **University of Information Technology,Yangon,Myanmar
soesoelwin@ucsy.edu.mm,khinthandarnwet@ucsy.edu.mm

Abstract

Due to an exponential growth in the generation of textual data, tools and mechanisms for automatic summarization of documents is needed. Text summarization is currently a major research topic in Natural Language Processing. There are various approaches to generate text summary. Among them, we proposed Myanmar text summarization using latent semantic analysis (LSA). Latent semantic analysis (LSA) is a technique in natural language processing, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA is a retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. There is no LSA based sentence extraction in Myanmar language. This is the first LSA based Text Summarizer in Myanmar. We summarize Myanmar news from Myanmar official websites such as 7day daily, new-eleven, ThithtooLwin, etc.

Keywords: *text summarization, LSA*

1. Introduction

With the explosion of data available on the Web in the form of unstructured text, efficient methods of summarizing text are becoming more important today. Text summarization is the creation of a shortened version of a document by a computer program. It extracts the most important points of the original text.

A text document generally consists of a set of topics. Topics are explained by some sentences and other sentences in the document are used to make the topics more readable and complete. The summary of a document should be able to cover all the topics presented in the original text and simultaneously should be able to keep the redundancy to a minimum level [1]. Summary should include important sentences that explains the topics covered in the original text.

Summarization system can be divided into two categories: extractive and abstractive summarization. Extractive summarization extracts salient words/sentences from documents and group them to produce summary without changing [10]. Abstractive summarization examines the source text and generate the concise summary that contains some novel sentences not present in original document.

Summarization methods can be categorized according to what they generate and how they generate it. A summary can be extracted from a single document or from multiple documents. If a summary is generated from a single document, it is known as single document summarization. On the other hand, if a single summary is generated from multiple documents on the same subject, this is known as multi -document summarization.

Summaries are also divided into generic summaries and query-based summaries. Generic summarization systems generate summaries containing main topics of documents. In query-based summarization, the generated summaries contain the sentences that are related to the given queries.

There are several research projects concerned with automatic text summarization for English and European languages and in Asian languages such as Chinese and Japanese. However, there is little ongoing research in Myanmar text summarization. In this paper, generic extractive Myanmar text summarization system based on LSA is proposed.

2. Related Works

Researchers have been working actively on text summarization within the Natural Language

Processing (NLP) to create a better and more efficient summary. The researchers have been developing for summarization using latent semantic analysis method to achieve better summary.

W.T.Kyaw[17] used CRF (Conditional Random Field) to summarize Myanmar disaster news that is based on Information Extraction. This paper proposed multi-documents, query-focused, extractive and informative Myanmar text Summarization framework. It is not sentence extraction. It is word level summary concerned with natural disasters such as date, time, Place. Our system extracts important sentence based on LSA.

M. T. Naing proposed Summary generation system is based on semantics roles. Automatic prominal anaphora resolution in Myanmar text is used for summary generation. In sematic role labeling, argument identification and argument classification are developed using Myanmar Verb Frame Resource. The system cannot identify verbs in sentences with two main verbs [18].

Gong and Liu [5] proposed summarization of news with the use of LSA as a way to identify the significant topics in the documents. SVD is applied to matrix A to decompose into three matrices as follows: A=U∑VT. They proposed that the row of the matrix VT can be considered as various topics covered in the original text. And finally, they reproduce each row of matrix VT successively and extract a sentence from it which has maximum values. They select one sentence for each topic according to topic importance.

M.G. Ozsoy, I. Cicekli and F.N. Ferda Nur Alpaslan [2] explained LSA-based summarization algorithms and evaluated on Turkish and English document. While creating summaries using LSA, there are various approaches for selection of sentences. Among them, they compared five algorithms: Gong and liu, Steinberger and Jezeck, Murry, cross method and topic method. Cross method and topic method are proposed by authors of that paper. They showed that among LSA-based approaches, the cross method performs better than other approaches. They observed that cross method does not perform well in shorter documents.

J.Steinberger and K.Jezek [6] described a generic text summarization method which used the latent semantic analysis technique to identify semantically important sentences. Summarization ratio is 20%. They suggested two new evaluation methods based on LSA, which measure content similarity between an original document and its summary. These two evaluation methods are: similarity of the main topic and similarity of the term significance.

O.foong, S.Yong and F.Jaid [7] investigated lsa based extractive text summarization in mobile android platform. The summary was compressed to 20% from the original document. The lsa produce the summary output with an average f-score of 0.36.

3. Latent Semantic Analysis (LSA)

LSA is an algebraic method, which can analyze relations between terms and sentences of a given set of documents. LSA uses context of the input document and extracts information such as which words are used together and which common words are seen in different sentences. High number of common words among sentences indicates that the sentences are semantically related [2]. It uses SVD (Singular Value decomposition) for decomposing matrices. SVD is a numerical process, which is often used for data reduction, but also for classification, searching in documents and for text summarization [3].

There are three main steps in Latent Semantic Analysis. These steps are as follows:

- Input Matrix Creation.
- Singular Value Decomposition.
- Sentence Selection.

3.1. Input Matrix Creation

The input document is represented in a matrix form to perform the calculations. A matrix is created which represents the input text. The row of the matrix represents the words in the sentences and column represents the sentences of the input document. The cells of matrix represent the importance of words in sentences. There are various methods to represent importance of words. These approaches are:

- ***Number of Occurrence***: The cell is filled with frequency of the word in the sentence.
- ***Binary Representation***: If a word occurs in the sentence the cell is filled 1, otherwise the cell value is 0.
- ***Root Type***: If the root type of the word is Noun, cell value is the frequency of the word, otherwise the cell value is 0.
- **Term Frequency-Inverse Document Frequency**:

$$TF=\frac{\text{Frequency of word i in sentence j}}{\text{Sum of frequencies of all words in sentence j}}$$

$$IDF=\log\left(\frac{\text{Number of sentences in input text}}{\text{Number of sentences containing word i}}\right)$$

- **Modified tf-idf:** Cell value are first calculated with tf-idf value. If cell value is less than or equal to average TF-IDF values in the associated row, then set them zero.

3.2. Singular Value Decomposition

LSA process the term-sentence matrix through the algorithm called Singular Value Decomposition (SVD) [8]. SVD is a method of word co-occurrence analysis. It is based on theorem from linear algebra. The input matrix A is decomposed into three matrices U, V, ∑. It is shown in equation (1).

$$A=U \sum V \quad (1)$$

U= word×concept matrix

∑=Scaling values, diagonal descending matrix

V= sentence×concept matrix

The algorithm of SVD:

1. Compute AA^T
2. Calculate the Eigen values and Eigen vectors of AA^T
3. Compute $A^T A$
4. Calculate the Eigen values and Eigen vectors of $A^T A$
5. Calculate the square root of the common positive Eigen values of AA^T and A^TA
6. Finally, assign the computed values to U, ∑, V^T.

3.3. Sentence selection

Using the results of SVD, different algorithms use different approaches to select important sentences. Different algorithms are proposed to select important sentences from the document for summarization using the results of SVD [2]. These algorithms are summarized in table 1.

Table1. Sentence Selection Approaches

Algorithms with LSA approach	Algorithm	Features for sentence selection
Gong and lius approach	Gong and Lius Method	**It is based on** 1.Matrix V^T
Steinberger and Iezek's Approach,	Lengthy Method	1. Matrix V^T 2. The length of the sentence vector
Murray, Renals and Carletta's approach	Murray, Renals and Carletta's Method	1. Matrix V^T 2. ∑ matrices
Ozsoy's approach	Cross Method	1.Matrix V^T 2.The average value of each sentence 3.The total length of each sentence vector
	Topic Method	1.Matrix V^T 2. The creation of concept x concept matrix 3.The strength values of each concept 4. Discovering the main-concepts and sub-concepts

4. Proposed System

This section presents our proposed Myanmar text summarization system. Figure 1 describes overall architecture of proposed system.

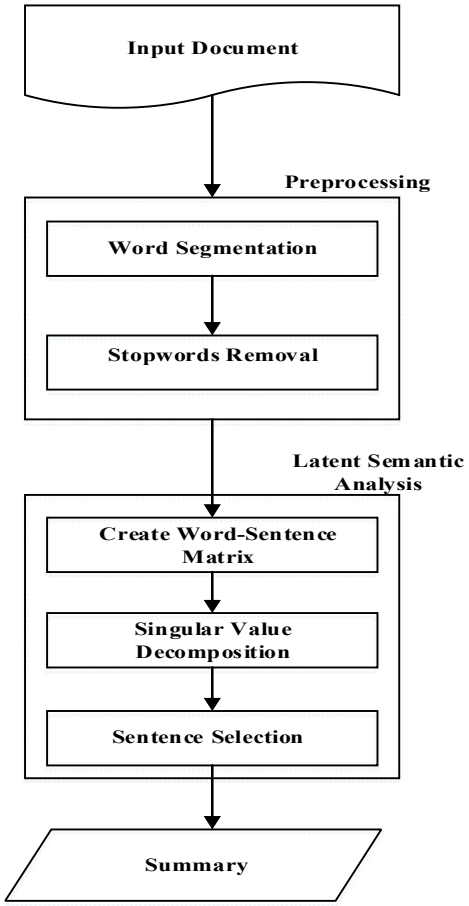

Figure 1. proposed system for text summarization

4.1. Word Segmentation

Word segmentation is the process of determining word boundaries in a text. In English language, words boundaries can be easily determined by space. In Myanmar language does not have white space between words.

The news data are converted to Unicode using Burmese font converter [11]. And then, we needed to perform word segmentation. In this paper, word segmentation is performed according to Burmese word segmentation program using Foma-generated Finite State Automata [9]. They used a foma regular expression and a readily available Burmese word list to create finite state automata to segment text into words using longest match with dictionary. Currently word lists have 41343 words. Local news contains the words such as ဒေါ်အောင်ဆန်းစုကြည်, ဘတ်ဂျတ်, ဖေဘွတ်ခ် ဘက်တီးရီးယား. But these words are not included in word list. The words such as အီရတ်, သီရိလင်္ကာ, ဗုံးကြဲ are included in international news but these words are missing in word list. Therefore, such words are added to word list. Now, word lists have 41543 words.

The performance and accuracy of word segmentation is very important because text summarization accuracy totally depends on preprocessing. The accuracy of Burmese word segmentation program using Foma-generated Finite State Automata is 86.92 %. The sample document after word segmentation is shown in table 2.

Table 2. Input Document after Word Segmentation

S1	အိန္ဒိယ|စစ်တပ်| ၏ |ရဟတ်ယာဉ် |တစ်စီး |ပျက်ကျ |ခဲ ၂ဦး |လိုက်ပါသွားသူ |၇ |ဦ |းစလုံး |သေဆုံး| ခဲ့ |ကြောင်း| သိရှိရသည်
S2	အိန္ဒိယ |လေတပ် |က |နိုင်ငံ |အရှေ့| မြောက်ပိုင်း| တရုတ် |နှင့် |နယ်နိမိတ် |အနီး |လူသူ |အရောက်အပေါက်| နည်းပါး ရာ |ဒေသ |တွင် |ရဟတ်ယာဉ် |ပျက်ကျ| ခဲ့| ခြင်းဟ ု |အတည်ပြုခဲ့သည်။
S3	ရဟတ်ယာဉ် |သည် |အဆိုပါ |ဒေသ ရှိ| အခြေစိုက် |စခန်း |တစ်ခု |မှ| ပျံတက် |လာ| ပြီး| မကြာမီ|တွင်| ပျက်ကျ| ခဲ့ခြင်း| ဖြစ်သည်|

4.2. Stopwords removal

Stopwords do not represent the main idea of the input text. Stopwords in the input document is removed. Examples of Stopwords are described in table 3.

Table3. Stopwords in Myanmar

Examples of Stopwords in Myanmar
ကြောင့်, က, နှင့်, ,၏ ,ဖြင့်,သို့ ,ရန်,ထို ,ယင်း ,သည်, မည်သည် ,မှာ, မှ ,မည် , အား ,တွင် ,ကို ,ပြီး ,ဖို့ လည်း, ထိ ,တိုင် ,စွာ ,ကား ,လုံး ,သော ,လျှင် ,များ,ခဲ့ , ကြောင်း, ဟု

4.3. Term-Sentence Matrix Creation

Input matrix is created based on the source document where columns represent sentence and rows represent terms. Frequency of the word in the sentence is filled in cell values of matrix. Input matrix is shown in figure 2.

	S1	S2	S3
အိန္ဒိယ	1	1	0
စစ်တပ်	1	0	0
ရဟတ်ယာဉ်	1	1	1
တစ်စီး	1	0	0
ပျက်ကျ	1	1	1
လိုက်ပါသွားသူ	1	0	0
ဂျီး	1	0	0
သေဆုံး	1	0	0
သိရှိ	1	0	0
လေတပ်	0	1	0
နိုင်ငံ	0	1	0
အရှေ့	0	1	0
မြောက်	0	1	0
တရုတ်	0	1	0
နယ်နိမိတ်	0	1	0
လူသူ	0	1	0
အရောက်အပေါက်	0	1	0
အတည်ပြု	0	1	0
ဒေသ	0	1	1
အခြေစိုက်စခန်း	0	0	1
ပျံတတ်	0	0	1
မကြာမီ	0	0	1

Figure 2. Term-sentence Matrix of document

4.4. Singular value decomposition

SVD is used to reduce dimension of term-by-document matrix. By applying SVD algorithm on input matrix, three matrices U, $\sum$, V^T are shown in figure 3,4, and 5.

$$U=\begin{bmatrix} -0.320 & 0.123 & 0.234 \\ -0.114 & 0.312 & 0.122 \\ -0.406 & 0.162 & -0.199 \\ -0.114 & 0.312 & 0.122 \\ -0.406 & 0.162 & -0.199 \\ -0.114 & 0.312 & 0.122 \\ -0.114 & 0.312 & 0.122 \\ -0.114 & 0.312 & 0.122 \\ -0.114 & 0.312 & 0.122 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.206 & -0.189 & 0.112 \\ -0.292 & -0.150 & -0.321 \\ -0.085 & 0.039 & -0.433 \\ -0.085 & 0.039 & -0.433 \\ -0.085 & 0.039 & -0.433 \end{bmatrix}$$

Figure 3. U word × concept matrix

$$\sum=\begin{bmatrix} 3.988 & 0 & 0 \\ 0 & 2.726 & 0 \\ 0 & 0 & 2.159 \end{bmatrix}$$

Figure 4. diagonal descending matrix ∑

$$V^T=\begin{bmatrix} 0.456 & -0.822 & -0.341 \\ 0.850 & -0.516 & 0.107 \\ 0.264 & 0.241 & -0.934 \end{bmatrix}$$

Figure 5. V^T concept × sentence matrix

To get U, ∑, V^T matrices, [4] is used. We used cross method to select sentence included in summary.

4.5. Sentence Selection

After creating input matrix and singular value decomposition of the matrix, sentence selection is performed to generate output summary. In this paper, cross method is used for sentence selection.

In cross method, V^T matrix and ∑ are used to select sentences. Firstly, V^T matrix is preprocessed to remove sentences that are not related to concepts. In table 4, rows represent concepts and column represents sentences. Average value of all concept are calculated. And then, if the cell value is less than or equal to average value, this cell value is set to be 0. V^T matrix after preprocessing is shown in table 4.

Table 4. V^T matrix after preprocessing

	sen1	sen2	sen3	Average
con1	-0.456	**0**	-0.341	-0.539
con2	0.85	**0**	**0**	0.147
con3	0.264	0.241	**0**	-0.143

And then, cell values are multiplied by with the values in ∑ matrix. The length of sentence vector is calculated by summing up all concepts in columns of V^T matrix, which represent the sentences. Length score calculation for each sentence is shown in table 5.

Table 5. length score calculation for each sentence

	Sen1	Sen2	Sen3
Con1	-1.81853	0	-0.736219
Con2	3.3898	0	0
Con3	1.05283	0.656966	0
length	**2.6241**	0.656966	-0.736219

In table 5, sen1 is the highest length score. Therefore, sentence 1 "အိန္ဒိယစစ်တပ် ၏ ရဟတ်ယာဉ်တစ်စီးပျက်ကျ ခဲ့ပြီး လိုက်ပါသွားသူ ၇ ဦး စလုံး သေဆုံးခဲ့ကြောင်း သိရှိရသည်။"is chosen as summary sentence. In this paper, average number of sentence in summarized document is 40% of the original document.

5. Experimental Work

Local news and international news are used for experiment. Summarization is done by Latent semantic analysis and performance measure is shown in this section.

5.1. Data used for Experiment

We collected Myanmar local news and international news from Myanmar news websites [12], [13], [14], [15], [16]. Now, we collected 55 local News and 76 international news from these websites. We will collect more data news in the future. Table 6 shows data set used for experiment.

Table 6. Data Set

	Local News	International News
Number of documents	55	76
Sentences per documents	275	456
Words per document	2860	4100

5.2. Performance Measure and Result

Human-generated summary is used as a reference to compare the summaries generated by machine for same document in this evaluation. Accuracy, precision and recall are used for evaluation parameter. Summarization accuracy of international news are less than local news because most of the words in international news are not contained in word list. Therefore, these words cannot be segmented correctly. Therefore, the accuracy of international news is less than local news. Table 7 describes evaluation measure formulas. And then performance measure by LSA is shown in table 8.

Table 7. Evaluation Measure

Accuracy	$\frac{tp + tn}{tp + fp + fn + tn}$
Precision	$\frac{tp}{tp + fp}$
Recall	$\frac{tp}{tp + fn}$

Where,

tp = True positive i.e. total number of correct sentences selected.

fp =False positive i.e. total number of sentences selected which are not correct.

fn = False negative i.e. total number of sentences which are correct but not selected.

tn= True negativei.e. total number of sentences which are not correct and also not selected.

Table 8. Performance Measure by LSA

Category	Accuracy	Precision	Recall
Local News	87%	60%	66%
International news	66%	50%	50%

6. Conclusion

Understanding a huge document without any abstract or summary is difficult and takes lot of time. This problem is solved with the automatic text summarizer. We proposed the Myanmar text Summarizer using latent semantic analysis model. This paper mainly concentrates on single document. In future work, we will combine graph based method to get better performance.

References

[1] P.V.Reddy, "Text Summarization Using Latent Semantic Analysis", International Journal of Technology and Engineering Science [IJTES], Volume 1[8], pp: 1309-1313, November 2013.

[2] M.G. Ozsoy, I. Cicekli, F.N. Ferda Nur Alpaslan, "Text Summarization of Turkish Texts using Latent Semantic Analysis", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–876, Beijing, August 2010.

[3] M. Campr, K. Ježek, "Comparative summarization via Latent Semantic Analysis", Department of Computer Science and Engineering University of West Bohemia.

[4] http://calculator.vhex.net/calculator/linear algebra/singular-value-decomposition.

[5] G.Yong, X.Liu, "Generic text summarization using relevance measure and latent semantic analysis". In: Proceedings of SIGIR'01 2001.

[6] J.Steinberger, K.Jezek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering.

[7] O.foong, S.Yong, F.Jaid, "Text Summarization using Latent Semantic Analysis Model in Mobile Android Platform", 2015 9th Asia Modelling Syposium".

[8] S.A.Babar, S.A.Thorat, "Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis", Computer Science & Engineering, Rajarambapu Institute of Technology, Sakharale, India, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 4 (May 2014).

[9] https://github.com/lwinmoe/segment.

[10] J. Kamala Geetha, N. Deepamala, "Kannada text summarization using Latent Semantic Analysis", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[11] http://www.myanmarengineer.org/converter/fontconv.htm

[12] http://www.moi.gov.mm/

[13] http://www.7daydaily.com/
[14] http://news-eleven.com/
[15] http://www.thithtoolwin.com/
[16] http://thevoicemyanmar.com/
[17] W.T.Kyaw, N.L.Thein and H.H.Htay, "Automatic Myanmar Text Summarization System", Proceeding of the 12th International Conference on Computer Applications (ICCA 2014),Yangon, Myanmar.
[18] M. T. Naing, A. Thida, "Automatic Myanmar Text Summarization System with Semantic roles", in the Proceedings of the 12th International Conference on Computer Applications (ICCA2014), Yangon, Myanmar, February2014, p. 217-223.