Single Document Text Summarization of a Resource-Poor Language using an Unsupervised Technique

Gunadeep Chetia, Gopal Chandra Hazarika

***Abstract:** Automatic text summarization of a resource-poor language is a challenging task. Unsupervised extractive techniques are often preferred for such languages due to scarcity of resources. Latent Semantic Analysis (LSA) is an unsupervised technique which automatically identifies semantically important sentences from a text document. Two methods based on Latent Semantic Analysis have been evaluated on two datasets of a resource-poor language using Singular Value Decomposition (SVD) on different vector-space models. The performance of the methods is evaluated using ROUGE-L scores obtained by comparing the system generated summaries with human generated model summaries. Both the methods are found to be performing better for shorter documents than longer ones.*

***Keywords:** Latent Semantic Analysis, Singular Value Decomposition, Text Summarization, Word-sentence Matrix.*

I. INTRODUCTION

Text summarization is a process of generating a summary of a document by presenting the important concepts of the original document. Summaries save a considerable amount of time of the reader and help in deciding whether to read a full document or not. They are also useful for many systems like Search Engine, Question Answering, Sentiment Analysis, News Aggregator etc. Considering the rapid increase in the amount of textual information on the Web, there is a growing need to produce summaries automatically through a computer. With the growing popularity of the Unicode standard in the World Wide Web, the textual contents of most of the non-English languages have also been increased on the web. Assamese is one such Indic language whose contents have been increased considerably in the last few years on the web. Especially the Assamese Wikipedia, online news portals, blogs, online magazines and social networking sites have contributed a lot in the increase of Assamese content on the web. But, like most other Indic languages, Assamese is a resource-poor language which lacks sufficient resources and tools required for advanced natural language processing tasks such as automatic summarization, machine translation etc. [1][2]. On the other hand, the rapid growth of Assamese content on the web and other digital platforms has necessitated the development of automatic text summarization system (ATS) for Assamese. Text summarization systems can be classified into two categories-extractive and abstractive [3]. Extractive summarization methods work by identifying salient sections of the text and then extracting those sections to form the summary. Typically, it involves ranking the sentences of the original text based on some criteria and extracting the higher ranked sentences for the summary. On the other hand, abstractive summarization works by analyzing the original text using some advanced NLP techniques and generating novel sentences to present the important concepts of the text in shorter form. Although summaries produced by human are often abstractive in nature, it is difficult to produce pure abstractive summaries through computer as it involves correct semantic representation, co-reference resolution and natural language generation [4]. In case of resource-poor language like Assamese, most of the research work focus on extractive summarization. No significant work has been found on the literature on abstractive summarization of Assamese text. Kalita et al. [5] proposed an extractive summarization method based on sentence similarity measure using Assamese Wordnet. In this paper, we present an unsupervised approach based on Latent Semantic Analysis (LSA) for extractive summarization of Assamese single-document text. We first perform the pre-processing steps on the input text by applying our own methods of stemming and lemmatization. Then we apply some existing techniques based on LSA on pre-processed Assamese text of various domains and analyze the performance of these techniques using ROUGE metrics. To the best of our knowledge, this is the first attempt to summarize Assamese text using LSA.

II. METHODOLOGY

Our approach mainly consists of two stages, viz. pre-processing and processing. During pre-processing phase we perform sentence segmentation, stop-word removal, lemmatization of the verbs and stemming of all other words in the input text. For stemming and lemmatization we use a hybrid approach proposed by us which is based on n-gram similarity matching technique [6].

Revised Manuscript Received on October 30, 2019.

* Correspondence Author
Gunadeep Chetia*, Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. Email: gunadeep@dibru.ac.in
Gopal Chandra Hazarika, Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. Email: gchazarika@gmail.com

After pre-processing we apply Latent Semantic Analysis (LSA) method to the normalized text. LSA is an algebraic-statistical method for extracting and representing the contextual meaning of words and similarity of sentences by statistical computations applied to an input text [7]. Gong and Liu [8] first published the idea of applying LSA to automatic text summarization. LSA uses Singular Value Decomposition (SVD), a mathematical matrix decomposition technique, for deriving the latent semantic structure from an input document. LSA method for summarization comprises three steps: creation of word-sentence matrix, applying SVD to the matrix, and selection of significant sentences.

A. Word-Sentence Matrix

To create the word-sentence matrix, we first remove the stop-words from the input document and then perform stemming and lemmatization. From the pre-processed input document, a matrix M is created by taking the words (stems or lemma) along the rows and sentences along the columns. For a text with *m* words and *n* sentences where without loss of generality $m>n$, it can be represented by $M=[M_{ij}]_{m\times n}$. The cell x_{ij} can be filled up with a value which represents the importance of the ith word in the jth sentence. Different vector space models can be used to fill up the cells[9]. We experiment with two such models described below:

(a) Binary feature model: This is a simple bag of words (BoW) model, where the presence of a word in a sentence is marked as a Boolean value. The cell value is filled with 1 if the word is present in the corresponding sentence and with 0 if it is absent.

(b) TF.IDF: The TF.IDF (Term Frequency.-Inverse Document frequency) score of a word represents the importance of a word in a document. If the word is frequent in the document but less frequent in other documents, then it can be said that the importance of the word is high in that document. Since we are dealing with single-document text summarization, we slightly modify this concept and instead of IDF we use Inverse Sentence Frequency (ISF), calculated as follows:

$$ISF(w) = \log \frac{n_w}{N} \qquad (1)$$

Where n_w is the number of sentences in which w occurs and N is the total number of sentences in the document. The TF part remains same, which is calculated as follows

$$TF(w) = f(w,s)/n_s \qquad (2)$$

Where f(w,s) is the number of times w appears in s and n_s is the total number of words in the sentence s. The value for each cell is obtained by multiplying the value obtained in equation (1) with that of equation (2).

B. Singular Value Decomposition

Singular Value Decomposition (SVD) is a matrix factorization method which decomposes a matrix into three other matrices. By applying SVD to our word-sentence matrix M, we decompose it into three matrices as follows:

$$M = U\,\Sigma\,V^T \qquad (3)$$

where U is an orthogonal matrix, Σ is a diagonal matrix and V is an orthogonal matrix. The columns of U and V are referred to as the left and right singular vectors, respectively, and the singular values of M are defined as the diagonal elements of Σ.

The three matrices are obtained through the following steps.

- Find MM^T
- Find the Eigen values and Eigenvectors of MM^T. The Eigenvectors of MM^T forms U after normalization
- Find M^TM.
- Find the Eigen values and Eigenvectors of M^TM. The Eigenvectors of M^TM forms V after normalization
- The square root of the non-zero and positive Eigen values of MM^T or M^TM sorted in descending order diagonally forms Σ.

This operation breaks down the original document into some linearly independent base vectors or concepts [10]. Thus the input matrix $M_{m\times n}$ gets decomposed into U which represents the words × extracted concepts matrix (m × n); ∑ represents the diagonal descending matrix (n × n) with singular values; and V is sentences × extracted concepts matrix (n × n). The significance of applying SVD is that it derives the hidden semantic structure from the document represented by matrix M in terms of some concepts [11]. In fact, the concepts represented by the singular vectors are obtained from the salient and recurring word combination patterns in the document. The corresponding singular values in ∑ indicate the degree of importance of the concepts within the document. The topmost sentences that best represent the concepts can be selected for summarization.

C. Selection of Sentences

Sentences representing the important concepts can be chosen by analyzing the decomposed matrices. We here experiment with two methods for sentence selection as follows: one proposed by Steinberger et al. and another proposed by Ozsoy et al.

Method-1: Here we make use of the V and ∑ matrices as proposed by Steinberger et al. in [12] for sentence selection. For each sentence vector in V, a score is calculated by computing the length of the sentence vector in n-dimensional concept space. The value of n is given to the method as a parameter. Mathematically, if v is the sentence vector in the concept space and σ_i is the singular value of the ith concept, the score of a sentence is given by

$$S(k) = \sqrt{\sum_{i=1}^{n} v_{k,i}^2 . \sigma_i^2} \qquad (4)$$

Basically, the score represents the extent to which the sentence explains a concept, weighted by the importance of that concept. The singular value represents the importance of a concept and it is used as a multiplication parameter to give more emphasis on the most important concepts. Thus the sentences are ranked based on the scores computed from the equation (4) and depending upon the compression percentage, some highly scored sentences are selected for the summary. In the resultant summary, the selected sentences are ordered as they appear in the original document to maintain the coherence.

Method-2: Here, input matrix creation and SVD steps are performed in the same way as describe in section II(A) and II(B) respectively, but sentence selection process is executed differently as proposed by Ozsoy et al [13]. Before executing the sentence selection process, a pre-processing step is performed on the rows of V^T matrix. The average score of each row in the V^T matrix is calculated and the cell values are set to zero if they are equal to or less than the average score. This is done in order to remove the overall effect of the sentences that are somehow associated with the concept, but not so significant for that concept [14].

III. SUMMARY EVALUATION

We have used two different datasets in Assamese for evaluating the summarization methods. The first dataset (Dataset A) contains 60 news documents collected from Assamese online news sites. The second dataset (Dataset B) contains 60 articles from different areas collected from Assamese e-magazines. Dataset B contains longer articles than Dataset A. Each document in Dataset A consists of 25-40 sentences, whereas each document in Dataset B consists of 70-100 sentences.

System generated summaries are usually evaluated by comparing against model summaries produced by human. A number of metrics exist in the literature for evaluating automatic summarizers [15]. ROUGE (Recall Oriented Understudy of Gisting Evaluation) is one of the popular methods for evaluating summarizers [16]. We use ROUGE-L metrics to compare our automatic summaries with model summaries. ROUGE-L is based on the idea of longest common subsequence (LCS) of two summaries [17]. If S and M are candidate and model summaries of lengths x and y respectively, then LCS-based F- measure score is calculated as follows:

$$P = \frac{LCS(S,M)}{x} \qquad (5)$$

$$R = \frac{LCS(S,M)}{y} \qquad (6)$$

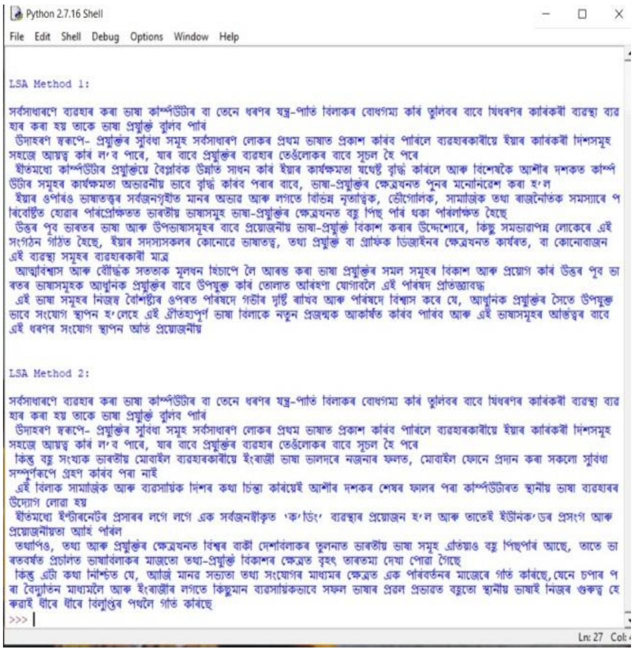
$$F = \frac{(1+\beta^2)RP}{R+\beta^2 P} \qquad (7)$$

where, LCS(S,M) is the longest common subsequence of S and M, and β=P/R. Equation (7) gives the ROUGE-L score of the two summaries.

Two individuals having knowledge of the Assamese language were assigned the job of generating the model summaries for the documents in both the datasets Dataset A and Dataset B. The summaries for the same documents were generated by the system using our proposed techniques. The length of the summary is fixed to be 30% of the original document. The model summaries are also generated at the same compression rate. After that, ROUGE-L scores are calculated for each pair of candidate summaries and model summaries.

IV. RESULTS AND DISCUSSION

Fig 1 shows an example of two automatic summaries generated using the two proposed methods on the same input document. In this example we applied the binary feature model for the word-sentence matrix creation. Fig 2 shows the example summaries of the same document generated using the same methods but considering the TF.IDF values while creating the word-sentence matrix.



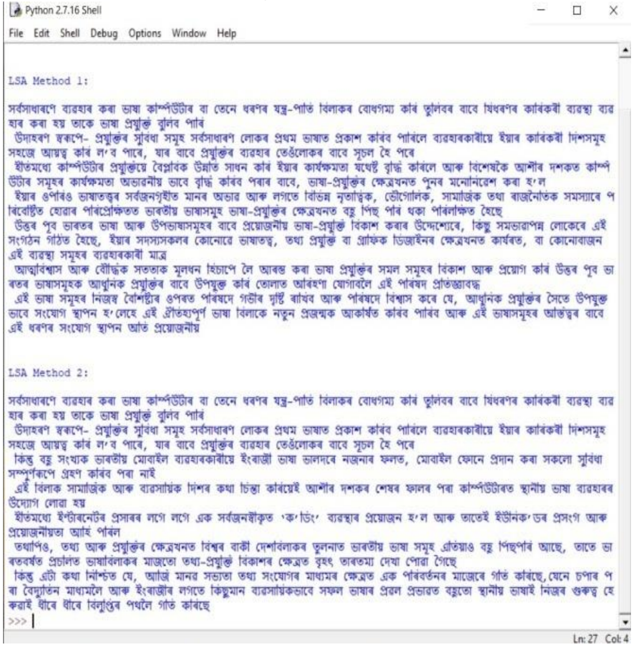
```
LSA Method 1:

সর্বসাধাৰণে ব্যৱহাৰ কৰা ভাষা কাম্পিউটাৰ বা তেনে ধৰণৰ যন্ত্ৰ-পাতি বিলাকৰ বোধগম্য কৰি তুলিবৰ বাবে যিধৰণৰ কাৰ্যকৰী ব্যৱস্থা ব্যৱ
হাৰ কৰা হয় তাকে ভাষা প্ৰযুক্তি বুলিব পাৰি
 উদাহৰণ স্বৰূপে- প্ৰযুক্তিৰ সুবিধা সমূহ সৰ্বসাধাৰণ লোকৰ প্ৰথম ভাষাত প্ৰকাশ কৰিব পাৰিলে ব্যৱহাৰকাৰীয়ে ইয়াৰ কাৰ্যকৰী দিশসমূহ
সহজে আয়ত্ব কৰি ল'ব পাৰে, যাৰ বাবে প্ৰযুক্তিৰ ব্যৱহাৰ তেওঁলোকৰ বাবে সুচল হৈ পৰে
 ইতিমধ্যে কাম্পিউটাৰ প্ৰযুক্তিয়ে বৈপ্লৱিক উন্নতি সাধন কৰি ইয়াৰ কাৰ্যক্ষমতা যথেষ্ট বৃদ্ধি কৰিলে আৰু বিশেষকৈ আশীৰ দশকত কাম্পি
উটাৰ সমূহৰ কাৰ্যক্ষমতা অভাৱনীয় ভাবে বৃদ্ধি কৰিব পৰাৰ বাবে, ভাষা-প্ৰযুক্তিৰ ক্ষেত্ৰখনত পুনৰ মনোনিৱেশ কৰা হ'ল
 ইয়াৰ ওপৰিও ভাষাতত্ত্বৰ সৰ্বজনগৃহীত মানৰ অভাৱ আৰু লগতে বিভিন্ন নৃতাত্ত্বিক, ভৌগোলিক, সামাজিক তথা ৰাজনৈতিক সমস্যাৰে প
ৰিবেষ্টিত হোৱাৰ পৰিপ্ৰেক্ষিতত ভাৰতীয় ভাষাসমূহ ভাষা-প্ৰযুক্তিৰ ক্ষেত্ৰখনত বহু পিছ পৰি থকা পৰিলক্ষিত হৈছে
 উত্তৰ পূব ভাৰতৰ ভাষা আৰু উপভাষাসমূহৰ বাবে প্ৰয়োজনীয় ভাষা-প্ৰযুক্তি বিকাশ কৰাৰ উদ্দেশ্যেৰে, কিছু সমভাৱাপন্ন লোকেৰে এই
সংগঠন গঠিত হৈছে, ইয়াৰ সদস্যসকলৰ কোনোৱে ভাষাতত্ব, তথ্য প্ৰযুক্তি বা গ্ৰাফিক ডিজাইনৰ ক্ষেত্ৰখনত কাৰ্যৰত, বা কোনোবাজন
এই ব্যৱস্থা সমূহৰ ব্যৱহাৰকাৰী মাত্ৰ
 আত্মবিশ্বাস আৰু বৌদ্ধিক সততাক মূলধন হিচাপে লৈ আৰম্ভ কৰা ভাষা প্ৰযুক্তিৰ সমল সমূহৰ বিকাশ আৰু প্ৰয়োগ কৰি উত্তৰ পূব ভা
ৰতৰ ভাষাসমূহক আধুনিক প্ৰযুক্তিৰ বাবে উপযুক্ত কৰি তোলাত অৰিহণা যোগাবলৈ এই পৰিষদ প্ৰতিজ্ঞাবদ্ধ
 এই ভাষা সমূহৰ নিজস্ব বৈশিষ্ট্যৰ ওপৰত পৰিষদে গভীৰ দৃষ্টি ৰাখিব আৰু পৰিষদে বিশ্বাস কৰে যে, আধুনিক প্ৰযুক্তিৰ সৈতে উপযুক্ত
ভাবে সংযোগ স্থাপন হ'লেহে এই ঐতিহ্যপূৰ্ণ ভাষা বিলাকে নতুন প্ৰজন্মক আকৰ্ষিত কৰিব পাৰিব আৰু এই ভাষাসমূহৰ অস্তিত্বৰ বাবে
এই ধৰণৰ সংযোগ স্থাপন অতি প্ৰয়োজনীয়


LSA Method 2:

সর্বসাধাৰণে ব্যৱহাৰ কৰা ভাষা কাম্পিউটাৰ বা তেনে ধৰণৰ যন্ত্ৰ-পাতি বিলাকৰ বোধগম্য কৰি তুলিবৰ বাবে যিধৰণৰ কাৰ্যকৰী ব্যৱস্থা ব্যৱ
হাৰ কৰা হয় তাকে ভাষা প্ৰযুক্তি বুলিব পাৰি
 উদাহৰণ স্বৰূপে- প্ৰযুক্তিৰ সুবিধা সমূহ সৰ্বসাধাৰণ লোকৰ প্ৰথম ভাষাত প্ৰকাশ কৰিব পাৰিলে ব্যৱহাৰকাৰীয়ে ইয়াৰ কাৰ্যকৰী দিশসমূহ
সহজে আয়ত্ব কৰি ল'ব পাৰে, যাৰ বাবে প্ৰযুক্তিৰ ব্যৱহাৰ তেওঁলোকৰ বাবে সুচল হৈ পৰে
 কিন্তু বহু সংখ্যক ভাৰতীয় মোবাইল ব্যৱহাৰকাৰীয়ে ইংৰাজী ভাষা ভালদৰে নজনাৰ ফলত, মোবাইল ফোনে প্ৰদান কৰা সকলো সুবিধা
সম্পূৰ্ণৰূপে গ্ৰহণ কৰিব পৰা নাই
 এই বিলাক সামাজিক আৰু ব্যৱসায়িক দিশৰ কথা চিন্তা কৰিয়েই আশীৰ দশকৰ শেষৰ ফালৰ পৰা কাম্পিউটাৰত স্থানীয় ভাষা ব্যৱহাৰৰ
উদ্যোগ লোৱা হয়
 ইতিমধ্যে ইণ্টাৰনেটৰ প্ৰসাৰৰ লগে লগে এক সৰ্বজনস্বীকৃত 'ক'ডিং' ব্যৱস্থাৰ প্ৰয়োজন হ'ল আৰু তাতেই ইউনিক'ডৰ প্ৰসংগ আৰু
প্ৰয়োজনীয়তা আহি পৰিল
 তথাপিও, তথ্য আৰু প্ৰযুক্তিৰ ক্ষেত্ৰখনত বিশ্বৰ বাকী দেশবিলাকৰ তুলনাত ভাৰতীয় ভাষা সমূহ এতিয়াও বহু পিছপৰি আছে, তাতে ভা
ৰতবৰ্ষত প্ৰচলিত ভাষাবিলাকৰ মাজতো তথ্য-প্ৰযুক্তি বিকাশৰ ক্ষেত্ৰত বৃহৎ তাৰতম্য দেখা পোৱা গৈছে
 কিন্তু এটা কথা নিশ্চিত যে, আজি মানৱ সভ্যতা তথ্য সংযোগৰ মাধ্যমৰ ক্ষেত্ৰত এক পৰিবৰ্তনৰ মাজেৰে গতি কৰিছে,যেনে চপাৰ প
ৰা বৈদ্যুতিন মাধ্যমলৈ আৰু ইংৰাজীৰ লগতে কিছুমান ব্যৱসায়িকভাবে সফল ভাষাৰ প্ৰবল প্ৰভাৱত বহুতো স্থানীয় ভাষাই নিজৰ গুৰুত্ব হে
ৰুৱাই ধীৰে ধীৰে বিলুপ্তিৰ পথলৈ গতি কৰিছে
>>> |
```

Fig 1: Example of two summaries generated using binary feature model



```
LSA Method 1:

সর্বসাধাৰণে ব্যৱহাৰ কৰা ভাষা কাম্পিউটাৰ বা তেনে ধৰণৰ যন্ত্ৰ-পাতি বিলাকৰ বোধগম্য কৰি তুলিবৰ বাবে যিধৰণৰ কাৰ্যকৰী ব্যৱস্থা ব্যৱ
হাৰ কৰা হয় তাকে ভাষা প্ৰযুক্তি বুলিব পাৰি
 উদাহৰণ স্বৰূপে- প্ৰযুক্তিৰ সুবিধা সমূহ সৰ্বসাধাৰণ লোকৰ প্ৰথম ভাষাত প্ৰকাশ কৰিব পাৰিলে ব্যৱহাৰকাৰীয়ে ইয়াৰ কাৰ্যকৰী দিশসমূহ
সহজে আয়ত্ব কৰি ল'ব পাৰে, যাৰ বাবে প্ৰযুক্তিৰ ব্যৱহাৰ তেওঁলোকৰ বাবে সুচল হৈ পৰে
 ইতিমধ্যে কাম্পিউটাৰ প্ৰযুক্তিয়ে বৈপ্লৱিক উন্নতি সাধন কৰি ইয়াৰ কাৰ্যক্ষমতা যথেষ্ট বৃদ্ধি কৰিলে আৰু বিশেষকৈ আশীৰ দশকত কাম্পি
উটাৰ সমূহৰ কাৰ্যক্ষমতা অভাৱনীয় ভাবে বৃদ্ধি কৰিব পৰাৰ বাবে, ভাষা-প্ৰযুক্তিৰ ক্ষেত্ৰখনত পুনৰ মনোনিৱেশ কৰা হ'ল
 ইয়াৰ ওপৰিও ভাষাতত্ত্বৰ সৰ্বজনগৃহীত মানৰ অভাৱ আৰু লগতে বিভিন্ন নৃতাত্ত্বিক, ভৌগোলিক, সামাজিক তথা ৰাজনৈতিক সমস্যাৰে প
ৰিবেষ্টিত হোৱাৰ পৰিপ্ৰেক্ষিতত ভাৰতীয় ভাষাসমূহ ভাষা-প্ৰযুক্তিৰ ক্ষেত্ৰখনত বহু পিছ পৰি থকা পৰিলক্ষিত হৈছে
 উত্তৰ পূব ভাৰতৰ ভাষা আৰু উপভাষাসমূহৰ বাবে প্ৰয়োজনীয় ভাষা-প্ৰযুক্তি বিকাশ কৰাৰ উদ্দেশ্যেৰে, কিছু সমভাৱাপন্ন লোকেৰে এই
সংগঠন গঠিত হৈছে, ইয়াৰ সদস্যসকলৰ কোনোৱে ভাষাতত্ব, তথ্য প্ৰযুক্তি বা গ্ৰাফিক ডিজাইনৰ ক্ষেত্ৰখনত কাৰ্যৰত, বা কোনোবাজন
এই ব্যৱস্থা সমূহৰ ব্যৱহাৰকাৰী মাত্ৰ
 আত্মবিশ্বাস আৰু বৌদ্ধিক সততাক মূলধন হিচাপে লৈ আৰম্ভ কৰা ভাষা প্ৰযুক্তিৰ সমল সমূহৰ বিকাশ আৰু প্ৰয়োগ কৰি উত্তৰ পূব ভা
ৰতৰ ভাষাসমূহক আধুনিক প্ৰযুক্তিৰ বাবে উপযুক্ত কৰি তোলাত অৰিহণা যোগাবলৈ এই পৰিষদ প্ৰতিজ্ঞাবদ্ধ
 এই ভাষা সমূহৰ নিজস্ব বৈশিষ্ট্যৰ ওপৰত পৰিষদে গভীৰ দৃষ্টি ৰাখিব আৰু পৰিষদে বিশ্বাস কৰে যে, আধুনিক প্ৰযুক্তিৰ সৈতে উপযুক্ত
ভাবে সংযোগ স্থাপন হ'লেহে এই ঐতিহ্যপূৰ্ণ ভাষা বিলাকে নতুন প্ৰজন্মক আকৰ্ষিত কৰিব পাৰিব আৰু এই ভাষাসমূহৰ অস্তিত্বৰ বাবে
এই ধৰণৰ সংযোগ স্থাপন অতি প্ৰয়োজনীয়


LSA Method 2:

সর্বসাধাৰণে ব্যৱহাৰ কৰা ভাষা কাম্পিউটাৰ বা তেনে ধৰণৰ যন্ত্ৰ-পাতি বিলাকৰ বোধগম্য কৰি তুলিবৰ বাবে যিধৰণৰ কাৰ্যকৰী ব্যৱস্থা ব্যৱ
হাৰ কৰা হয় তাকে ভাষা প্ৰযুক্তি বুলিব পাৰি
 উদাহৰণ স্বৰূপে- প্ৰযুক্তিৰ সুবিধা সমূহ সৰ্বসাধাৰণ লোকৰ প্ৰথম ভাষাত প্ৰকাশ কৰিব পাৰিলে ব্যৱহাৰকাৰীয়ে ইয়াৰ কাৰ্যকৰী দিশসমূহ
সহজে আয়ত্ব কৰি ল'ব পাৰে, যাৰ বাবে প্ৰযুক্তিৰ ব্যৱহাৰ তেওঁলোকৰ বাবে সুচল হৈ পৰে
 কিন্তু বহু সংখ্যক ভাৰতীয় মোবাইল ব্যৱহাৰকাৰীয়ে ইংৰাজী ভাষা ভালদৰে নজনাৰ ফলত, মোবাইল ফোনে প্ৰদান কৰা সকলো সুবিধা
সম্পূৰ্ণৰূপে গ্ৰহণ কৰিব পৰা নাই
 এই বিলাক সামাজিক আৰু ব্যৱসায়িক দিশৰ কথা চিন্তা কৰিয়েই আশীৰ দশকৰ শেষৰ ফালৰ পৰা কাম্পিউটাৰত স্থানীয় ভাষা ব্যৱহাৰৰ
উদ্যোগ লোৱা হয়
 ইতিমধ্যে ইণ্টাৰনেটৰ প্ৰসাৰৰ লগে লগে এক সৰ্বজনস্বীকৃত 'ক'ডিং' ব্যৱস্থাৰ প্ৰয়োজন হ'ল আৰু তাতেই ইউনিক'ডৰ প্ৰসংগ আৰু
প্ৰয়োজনীয়তা আহি পৰিল
 তথাপিও, তথ্য আৰু প্ৰযুক্তিৰ ক্ষেত্ৰখনত বিশ্বৰ বাকী দেশবিলাকৰ তুলনাত ভাৰতীয় ভাষা সমূহ এতিয়াও বহু পিছপৰি আছে, তাতে ভা
ৰতবৰ্ষত প্ৰচলিত ভাষাবিলাকৰ মাজতো তথ্য-প্ৰযুক্তি বিকাশৰ ক্ষেত্ৰত বৃহৎ তাৰতম্য দেখা পোৱা গৈছে
 কিন্তু এটা কথা নিশ্চিত যে, আজি মানৱ সভ্যতা তথ্য সংযোগৰ মাধ্যমৰ ক্ষেত্ৰত এক পৰিবৰ্তনৰ মাজেৰে গতি কৰিছে,যেনে চপাৰ প
ৰা বৈদ্যুতিন মাধ্যমলৈ আৰু ইংৰাজীৰ লগতে কিছুমান ব্যৱসায়িকভাবে সফল ভাষাৰ প্ৰবল প্ৰভাৱত বহুতো স্থানীয় ভাষাই নিজৰ গুৰুত্ব হে
ৰুৱাই ধীৰে ধীৰে বিলুপ্তিৰ পথলৈ গতি কৰিছে
>>> |
```

Fig 2: Example of two summaries generated using TF.IDF model

Table 1 and Table 2 shows the average ROUGE-L scores for all the documents in datasets Dataset A and Dataset B respectively.

Table- I: Average ROUGE-L Scores for Dataset A

		Summarization Method	
		Method 1	Method 2
Matrix Creation Model	Binary	0.74822	0.75187
	TF.IDF	0.69202	0.74887

Table- II: Average ROUGE-L Scores for Dataset B

		Summarization Method	
		Method 1	Method 2
Matrix Creation Model	Binary	0.56386	0.57948
	TF.IDF	0.48146	0.57148

From the ROUGE-L results it is observed that method 2 produced better results than method 1 for both the datasets. The binary model for input matrix creation also produced better results than the TF.IDF model for both the datasets. However, method 2 is not much affected by the matrix creation model. The poor performance of TF.IDF model can be explained by the fact that sometimes significant words indicative of a concept occur frequently in the document resulting in low IDF (or ISF) value. On the other hand some non-significant words may occur very infrequently along the document and thereby increasing the ISF value.

It is also observed that ROUGE-L scores for Dataset A are significantly higher than those of Dataset B. This can be attributed to the fact that Dataset A contains shorter documents than Dataset B and the number of options for the human summary generator will also be less. Hence, the difference between the system generated summaries and model summaries will not be very high.

V. CONCLUSION

In this paper we have presented two methods based on Latent Semantic Analysis for single document extractive summarization. We have evaluated the methods using two different datasets that are in Assamese. The comparison of these methods was done by using ROUGE-L F measure scores. We have observed that binary feature model of word-sentence matrix creation technique performs better than the TF.IDF model. We have also observed that method 2 which is based on the sentence selection techniques proposed by Ozsoy et al. gives better results in our datasets than the method 1 based on the sentence selection techniques proposed by Steinberger et al. One limitation of LSA is that its performance decreases for documents with large number of sentences. But, still we have seen that the results are satisfactory for a resource-poor language like Assamese. Since LSA does not involve much linguistic work except in the pre-processing phase, it can be conveniently applied to extractive text summarization of other resource-poor languages.

REFERENCES

1. Islam, S., Devi, M. I., & Purkayastha, B. S. (2017). A study on various applications of NLP developed for North-East languages. International Journal on Computer Science and Engineering, 9(6), 368-378.
2. Sreelekha, S., & Bhattacharyya, P. (2019). Indowordnet's help in Indian language machine translation. AI & SOCIETY. doi:10.1007/s00146-019-00907-w
3. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., D., E., B., J., & Kochut, K. (2017). Text summarization techniques: A brief survey. International Journal of Advanced Computer Science and Applications, 8(10). doi:10.14569/ijacsa.2017.081052
4. Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457-479. doi:10.1613/jair.1523
5. Kalita, C., Saharia, N., & Sharma, U. (2012). An extractive approach of text summarization of Assamese using WordNet. In Global WordNet Conference (GWC-12).
6. Chetia, G., & Hazarika, G. C. (2018). Pre-processing phase of automatic text summarization for the Assamese language. International Journal of Computer Sciences and Engineering, 6(10), 159-163. doi:10.26438/ijcse/v6i10.159163
7. Wang, Y., & Ma, J. (2013). A comprehensive method for text summarization based on Latent Semantic Analysis. Communications in Computer and Information Science, 394-401. doi:10.1007/978-3-642-41644-6_38
8. Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01. doi:10.1145/383952.383955
9. Babar, S. A., & Patil, P. D. (2015). Improving performance of text summarization. Procedia Computer Science, 46, 354-363.
10. Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM review, 37(4), 573-595.
11. Steinberger, J., & Ježek, K. (2004). Text summarization and singular value decomposition. In International Conference on Advances in Information Systems (pp. 245-254). Springer, Berlin, Heidelberg.
12. Steinberger, J., & Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. Proc. ISIM, 4, 93-100.
13. Ozsoy, M. G., Cicekli, I., & Alpaslan, F. N. (2010). Text summarization of turkish texts using latent semantic analysis. In Proceedings of the 23rd international conference on computational linguistics (pp. 869-876). Association for Computational Linguistics.
14. Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using latent semantic analysis. Journal of Information Science, 37(4), 405-417.
15. Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1), 1-66.
16. Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
17. Cohan, A., & Goharian, N. (2016). Revisiting summarization evaluation for scientific articles. arXiv preprint arXiv:1604.00400.

AUTHORS PROFILE

Gunadeep Chetia is currently pursuing Ph. D. at Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. He has more than six years of teaching experience. He has also developed a number of tools related to language technology. His area of research is Natural Language Processing.

Gopal Chandra Hazarika is currently working as the Chairperson of Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. He has more than thirty years of teaching and research experience. He has published a number of research papers on Mathematics and Computer Science in reputed international journals.