TEXT ANALYSIS FOR BENGALI LONG TEXT SUMMARIZATION USING DEEP LEARNING

BY
MD. MAHIBULLA HASAN
ID: 171-15-9411
&
MD. SAYDUR RAHMAN
ID: 171-15-8853

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

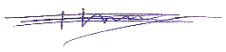
Mr. Sheikh Abujar
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 28, 2021

APPROVAL

This Project titled "**Text analysis for Bengali long text summarization using deep learning**", submitted by Md. Mahibulla Hasan, ID: 171-15-9411 and Md. Saydur Rahman, ID: 171-15-8853 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on January 28, 2021.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Jueal Mia
Senior Lecturer — **Internal Examiner**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Sheikh Abujar, Sr. Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Sheikh Abujar
Sr. Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

(Md. Mahibulla Hasan)
ID: 171-15-9411
Department of Computer Science and Engineering
Daffodil International University

(Md. Saydur Rahman)
ID: 171-15-8853
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

We have been trying our best effort to this thesis. It was not imaginable without the caring sustenance and support individuals with each other. We would like to express our sincere gratefulness to all those who provided us the opportunity to complete this report.

At first, we express our sincerest thanks and appreciation to the almighty Allah for His heavenly blessings which allowed us to complete this thesis successfully.

A special thanks to our parents for bearing us on this earth and admiration for endless support.

Also gratefulness to our supervisor, Mr. Sheikh Abujar, Sr.Lecturer of CSE Department, whose influence in inspiring propositions and reassurance, assisted us to coordinate our thesis specifically in writing this report. His limitless endurance, academic supervision, endless and bouncing direction, productive criticism, valuable instruction has made it imaginable to fulfill this thesis.

Moreover, we would like to concede with much gratitude the vital role of our department head, Professor Dr. Touhid Bhuiyan, who provided us with his expensive time and caring support to finish this thesis. We also give our sincere thanks to all the faculty members.

ABSTRACT

In the text summaries, a description is the system of the scale of one or more documents & the key detail of the document is provided by the curtail passage. In this present era of information technology, we always rely on online information that is raised in need of a summary of the original text. There is a lot of methods already implemented for other languages like English text summarize but the effort is now underway for Bengali text summarization. In this paper, we will propose RNN's (Recurrent Neural Network) deep learning model based extract text abbreviation strategy. So the categorization strategy is useful for description or not. For the back extension process, we have used Long Short-Term Memory(LSTM) and Gated Recurrent Units (GRU) in this article. But we used Long Short-Term Memory (LSTM) so it's more positive than that, and in terms of Rouge-I, Rouge-II and Rouge-III, we realize the typical F-1 scores being 0.65, 0.61, 0.58.

Keywords: Supervised learning, Sequence classification, Deep neural network, Extractive summary, Bengali text.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

There are huge data accessible internet also growing quickly. People searching their desired information in every moment. Sometimes they find what they exactly want and sometimes not also they are overcome in similar data. For a hunter's chosen records,

It is also intense to have real time to browse through all the attached documents presented on the internet. In these types of cases, automated text summarization can be very effective. The description of a large text should be read by everyone to make sure it is real for them. It has recently developed an important component for productivity indexing, tracking of forecasts, comparisons to news and blogs. Computer-generated summaries are bias-free and capable of collecting vast amounts of appropriate details. Automated text abbreviations often result from data loss but learning large amounts of data has become important.

Based on the procedure, different types of summary symptoms may be generated. Two types of input documents can depend on this. It may be poorly classified as a text summary for a single document because a summary originates from a single document. If the input is two or more documents of the same type, it is called a text summary of multiple documents.

By its essence, the abstract may also be called since a summary extract is generated by sketching the most relevant details from an unchanging text. The abstract description, on the other hand, introduces an article's core concept by mimicking the relevant portion of the article. An excerpt summary does not contain any text that is not part of a single document, although the author of the abstract summary may apply new words or phrases to the related text summary.Thus, extractive abbreviations are simpler to add than abstract abbreviations in most situations.

The two forms can be distinguished as index and informative summaries based on the content of the summary of the article. An index summary summarizes the initial concept of the whole text without explaining the content in the text.

In this paper we focus on summarizing the text of a single document. The sentence of every document designed vector score 1 is remain in the sentence and absence 0.

1.2 Motivation

The main objective of this paper is to introduce a compressed form of an article using deep learning method. RNN is a powerful model which operates very efficiently on text sequence. Several extractive and abstractive text summarization approach have been carried out on Bengali text throughout recent years as far our knowledge.

Many methods have been conducted for Bengali text summarization such as word scoring, sentence ranking, Graph scoring [8], cluster based method, TF- IDF method. On the best of our knowledge no neural network based approach was conducted for Bengali text document. There are approximately 150 national, regional and online Bengali newspapers in Bangladesh [9]. Most of them are covering various news everyday which is producing a large amount of data carrying on same topic. These data require summarization for future reader and analyzer.

We have followed a simple approach which is known as sequence classification. The work presented in [7], motivated us to conduct our summarization process based on neural network. In their work the sentences were labeled as crucial or not for the summary. Their approach indicated that the trained summary is intended to learn the pattern which helps in contributing summary. Whenever a new article passed on, the trained pattern will classify the sentences by giving score 0 to 1 which will produce a summary of certain sentences.

1.3 Rationale of the Study

There are a good number of researches going on English text summarization and the result are getting higher time by time. English text summarization is very much on the similar stage as human generated summarization. Bengali texts summarization is way backward compared to English text summarization as it faces some challenges of proper

linguistic tools, efficient stemming and lemmatizing, miscellaneous structure of Bengali text, unable to understand the context etc.

1.4 Outcome

Our research work aims at

1. Classify each sentence of the newspaper article whether they are part of the summary or not.
2. Generate summary which are nearly comparable to human generated summary.
3. Comparing the performance of our model with other previously recognized Bengali text summarization approach.

1.5 Report Layout

We addressed the implementation of inspiration, automated text description, the rationale of the analysis and the result of the research in this section. Later, the style of the study followed. In chapter 2, we will discuss about the background of our research topic. In chapter 3, we will discuss about the methodologies employed in our study. In chapter 4, we will discuss about the obtained results and discussion.

In chapter 5, we will discuss about the conclusion and future works.

CHAPTER 2
BACKGROUND

2.1 Introduction

Text summarization is one of most valuable parts of Natural Language processing. The first approach of text summarization was introduced in 1950. Since then several method was evaluated and suggested. Earlier on some simple techniques like position of word or sentences, frequency of words, terms from user queries and key phrases [1] are used. Bengali text summarization approach is also affected by those techniques.

2.2 Related Works

"The very first work of automatic text summarization was carried out by Luhn [10] in 1958 based on term frequency and the approach was extended by Baxendale [11] by comprising the cue words and position of the sentences in the document. These valuable contributions laid the foundation of computerized text summarization and from then researchers are eagerly contributing in this arena of Natural Language Processing.

In the paper of Sarker [4] as usual preprocessing and stemming are performed at first, all the sentences are ranked based on features like thematic terms, positional value and the length of the sentences. Thematic terms were considered if the TFIDF value of a term was greater than a predefined threshold. The top ranked k sentences were considered as desired summary.

A neural attention architecture was proposed by cheng and Lapata [12] for extracting words and sentences. Their encoder aims to deduct the variant of neural attention of the input article as uninterrupted sentence features and decoder extracts sentences based on the applied attention.

A sequence classification based Neural network model is also proposed by Nallapati et al. [13]. They have treated the sentences of the document as binary form depending upon their existence in the summary which is very similar to our proposed method. They have used GRU-RNN based neural network model and we found LSTM-RNN more favorable.

Some enhanced features can be found on the proposed approach of Verma and Nidhi [3]. Extra features like number of proper nouns, number of numerals.

efficient abstract summarization. Highest TF-IDF scored sentence is considered as the centroid sentence and then cosine similarity between the highest TF-ISF scored and every sentence is calculated which is termed as Sentence to cosine similarity. The produced feature matrix was used as output to a two layers Restricted Boltzmann Machine (RBM) hence an enhanced feature vector is produced. The enhanced feature vector values are added to produce a score for every sentence. Then the sentences are sorted in a descending order and the most efficient sentences was selected for the summary.

A Recursive Neural Network application for multi document summarization has come out from the approach of Cao et al. [14].They have developed a hierarchical regression process for the sentence ranking task. They have conducted their research on multi document summarization datasets the DUC 2001, 2002 and 2004 and showed that their proposed method exceeds R2N2 state-of-the-art extractive summarization approaches.

Akter et al. [8] proposed a summarization approach for Bengali single and multiple text document. They have used K means clustering method for the candidate summary. The centroids of the clusters are considered by the highest scored sentences. Sentence is scored by the TF-IDF value of each word. If any cue word is detected in the sentences, then the score of the sentence is increased by 1.

A document classification task is also inaugurated by Isonuma et. al. [15] for single document summary. They have evaluated their neural network based model on the documents of two financial based news publisher. Convolutional Neural Network(CNN) is used for sentence embedding from word embedding because of its efficiency on sentence level classification problem. Another neural network based model LSTM-RNN is used for extracting summaries from the document.

Topic based opinion summary for Bengali document is carried out by Das and Bandyopadhyay [16]. For distinguishing the sentiment information, they have used an annotation tool which annotate sentence for summary by pointing out the root words. Annotator spots the sentiment words according to their Part of Speech (POS) categories. K-means clustering is used for combining topic- sentiment. Finally, for Selecting the sentence of summary theme based relational graph is used and page rank algorithm is used for information recovery. Their approach is efficient for theme detection"[2].

2.3 Research Summary

A lot of contribution have been made for English text summarization. Many rule based and machine learning algorithm based Extractive and abstractive text summarization has been implemented for English texts. Deep learning models are the new player in this area. More accurate results have been achieved by using these deep learning models. For Bengali text summarization, all the efforts were rule based up to now. Extractive and abstractive text summarization for Bengali documents have been implemented throughout the recent years but they are not effective so far. We tried to implement Bengali text summarization in an extractive way.

2.4 Scope of the Problem

No deep learning based approach was conducted for Bengali text summarization as far as our knowledge. We have tried our best for summarizing Bengali single document text summarization using deep learning. Machine generated English text summarization has already reached to human generated text summarization. Due to difficulty of processing the Bengali text, the summarization of Bengali text did not reach that level yet. Efficient stemming and lemmatizing will be a revolution for Bengali text summarization problem.

2.5 Challenges

The main problem we faced for this work is limitations of data. Bengali data is very hard to collect. Processing the Bengali texts is another challenging matter for us. There is no rich annotated Bengali text summarization corpus. Besides LSTM as well as deep learning models need high specifications of hardware components. These models also require a large amount of time to operate.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Data Collection

Bengali data analysis is very challenging due to unavailability collecting dataset.There are 200 Bengali news article dataset used here. 3 sets are used [17] but it was arbitrary topic[17].We make a well-matched proposed model and need the best modelof each document so manually sort out best summary form the 3 set.Then from all the article consistent summaries save separate file as .txt file with UTF -8 encoding final dataset in Table 3.1 showing an example of collected dataset.

Table 3.1 Example of the collected data

Article: সুদীপ্ত দত্ত অর্জুন নামে এক কিশোরের লাশ আজ মঙ্গলবার সকালে ধানমন্ডি লেক থেকে উদ্ধার করেছে পুলিশ। সুদীপ্ত রাজধানীর রাইফেলস স্কুল অ্যান্ড কলেজের শিক্ষার্থী। সে এ বছরের এইচএসসি পরীক্ষার্থী। আজ সকাল সাড়ে আটটার দিকে ধানমন্ডি লেকে একজনের লাশ ভাসতে দেখে পুলিশে খবর দেয় সেখানে ঘুরতে আসা কয়েকজন ব্যক্তি। এরপর পুলিশ গিয়ে লাশ উদ্ধার করে। অর্জুন রায়ের বাজারে তার পরিবারের সঙ্গে থাকত। ধানমন্ডি থানার ভারপ্রাপ্ত কর্মকর্তা (ওসি-তদন্ত) হেলালউদ্দিন প্রথম আলোকে বলেন, ছেলেটির স্বাস্থ্যগত কিছু সমস্যা ছিল। তার মুখ থেকে দুর্গন্ধ বের হতো। এ নিয়ে অনেকে অনেক কথা বলতো। তা ছাড়া বাসা থেকে পড়াশোনার জন্য খুব চাপ দেওয়া হতো। পরিবারের ওপর রাগ করে গত ২৮ ফেব্রুয়ারি সে বাসা থেকে বের হয়ে যায়। এরপর তার পরিবারের পক্ষ থেকে থানায় একটি সাধারণ ডায়েরি (জিডি) করা হয়। সেখানে রাগ করে বাসা থেকে বের হওয়ার কথার উল্লেখ করা হয়েছে। ওই কর্মকর্তা জানান, পুলিশ লাশ এনে থানায় রাখে। এরপর ছেলেটির মা ছন্দা দত্ত ও বোন এসে লাশ শনাক্ত করে। ময়নাতদন্তের জন্য লাশ ঢাকা মেডিকেল কলেজে পাঠানো হয়েছে বলে জানান তিনি। দুই ভাই বোনের মধ্যে অর্জুন ছোট।
Summary: সুদীপ্ত দত্ত অর্জুন নামে এক কিশোরের লাশ আজ মঙ্গলবার সকালে ধানমন্ডি লেক থেকে উদ্ধার করেছে পুলিশ পরিবারের ওপর রাগ করে গত ২৮ ফেব্রুয়ারি সে বাসা থেকে বের হয়ে যায় এরপর তার পরিবারের পক্ষ থেকে থানায় একটি সাধারণ ডায়েরি (জিডি) করা হয় ওই কর্মকর্তা জানান, পুলিশ লাশ এনে থানায় রাখে এরপর ছেলেটির মা ছন্দা দত্ত ও বোন এসে লাশ শনাক্ত করে ময়নাতদন্তের জন্য লাশ ঢাকা মেডিকেল কলেজে পাঠানো হয়েছে বলে জানান তিনি ।

3.2 Dataset Creation

Our description model is derived by the root of supervised learning exection, for that we need an article that can mark each phrase as 0 or 1. Here, 0 means the sentence is not belong the summery and 1 means sentence belong in summary.

In order to do so, we need to split all the sentences from the article and the summaries. Then the documents with the labeling form are translated relative to the accompanying summaries of sentences.

With the aid of jellyfish 0.8.2, we went through this modification by using the python software. We also saved all the labeled article as .csv file with UTF -8 encoding further processing. Represented some labeled sentence in Table 3.2.

Table 3.2 Example of the labeled data

Sentence	Label
সুদীপ্ত দত্ত অর্জুন নামে এক কিশোরের লাশ আজ মঙ্গলবার সকালে ধানমন্ডি লেক থেকে উদ্ধার করেছে পুলিশ	1
এরপর তার পরিবারের পক্ষ থেকে থানায় একটি সাধারণ ডায়েরি (জিডি) করা হয়	0
এরপর ছেলেটির মা ছন্দা দত্ত ও বোন এসে লাশ শনাক্ত করে	0

The workflow of dataset creation is depicted in Figure 3.1 where the inputs are unannotated documents and summaries and outputs are labeled articles.

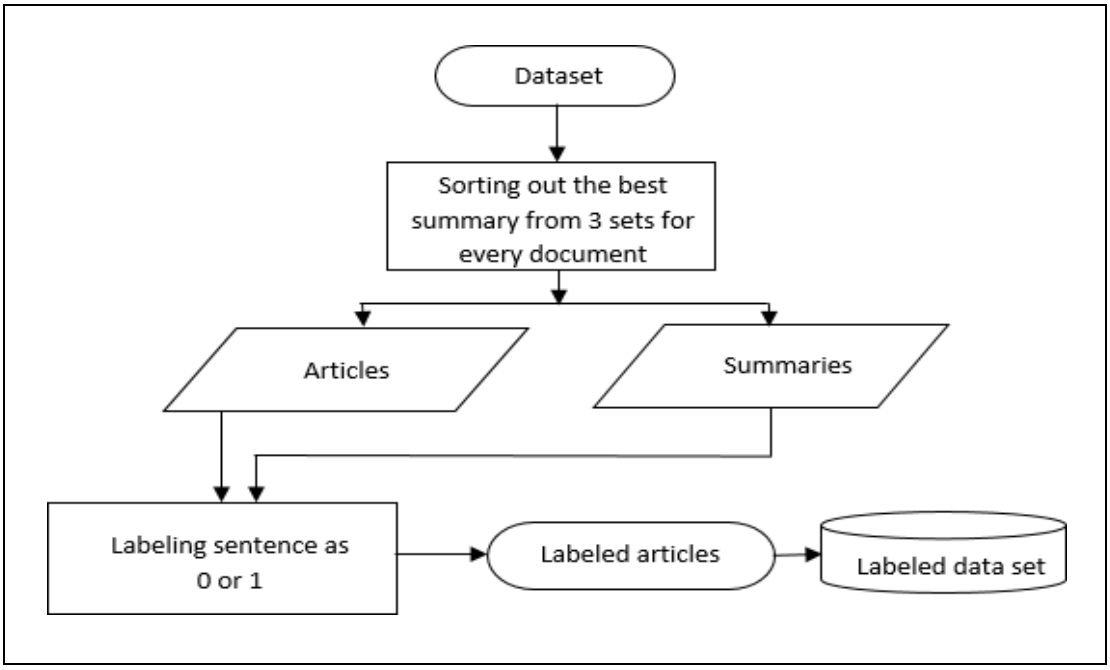


Figure 3.1 Dataset creation work flow

3.3 Proposed Model

In this working purpose, We used a model for summarizing the long Bengali single record by grouping of sequences. All the sentence of documents making binary classified subsequently visited all the sentence. Because binary classification ensure that the sentence remains the result of summary is or isn't.

The base of this model is LSTM full formed is Long Short-term Memory, building of deep neural network. The LSTM network attained very high performances along with solving slope disappearing or slope ignition problems [19]. Along with a single memory unit, the LSTM architecture has three entries in each cell.

The calculation of the mathematical operation of each cell can basically be represented as shown below:

$$K = [h_{t-1}, k_t] \quad \text{.....(1)}$$

$$o_t = \sigma(w_o \cdot K + b_o) \quad \text{..... (2)}$$

$$f_t = \sigma(w_f \cdot K + b_f) \quad \text{..... (3)}$$

$$i_t = \sigma(w_i \cdot K + b_i) \quad \text{..... (4)}$$

$$h_t = o_t * \tanh(c_t) \quad \text{.....(5)}$$

$$c_t = f_t * c_{t-1} + i_t * tanh(w_c \cdot K + b_c) \quad \text{.....(6)}$$

Here, w is the weighted matrices, $f \rightarrow$ forget gate , $o \rightarrow$ output gate and $i \rightarrow$ input gate. k_t present time step of input. b stands for biases and c_t means cell state. σ represents sigmoid function. Our model is organized with several sequential layer as Embedding layer, LSTM layer and Dense layer.In every sentence should be taken as a sequence of words and sequences of words will be converted as word vector by the current embedding of Embedding layer. The Embedding layer gives the output to the next LSTM layer of fixed length vector encoded sequences. The LSTM layer begins the calculation process through three gates and memory cells for each sequential input in each cell. It provides output to the next-Dense layer.

Outputs from the earlier layer are altered by the dense layer, based on the binary cataloging sigmoid activation function. The model validation and preparation forms a parameter and thus tests the weight of phrases during testing period to estimate the attachment in description.

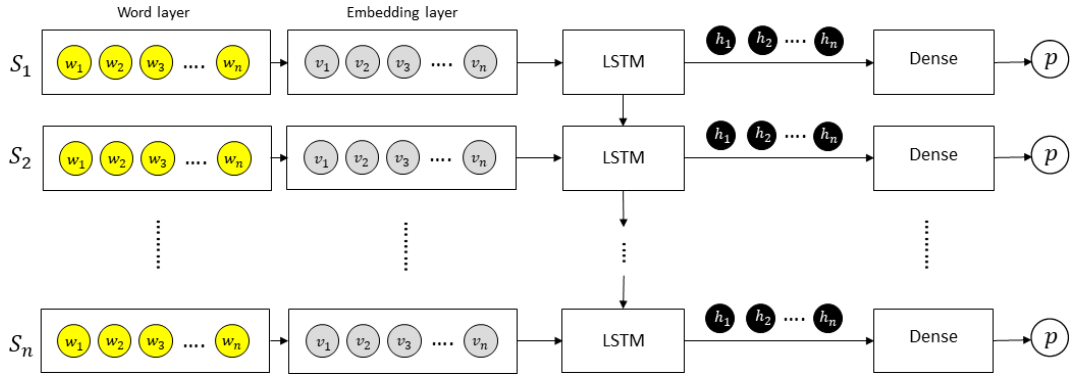


Figure 3.2 Our Model Architecture

Figure 3.4 reflects our model where w defines sentence terms, v defines each sentence's word vectors, and h stands for hidden vectors. The likelihood of any word is p.

3.4 Preprocessing

It should be correctly processed for more detailed outcomes in order to interpret textual data in machine learning. In the initial state, our data was very noisy with punctuation marks, numeric values and a large number of stop terms. The existence of certain forms of noise confuses the decision-making model. We removed the exclamation points from all of the phrases in our data set in the first instance. Then, after tokenization, numeric and non-alphabetic characters were omitted and stop words were eventually removed from [20]. Figure 3.3 illustrates a sentence preprocessing study.

ঢাকা-চট্টগ্রাম মহাসড়কের ওপর মঞ্চ বানিয়ে গতকাল বুধবার সকালে সভা করেছে নারায়ণগঞ্জের সোনারগাঁ উপজেলা যুবলীগ(উয়ু) এর ফলে গুরুত্বপূর্ণ মহাসড়কটিতে ১২ কিলোমিটারজুড়ে যানবাহনের সারি তৈরি হয় গতকাল সকাল ৭.৩০ মঞ্চের সামনে মহাসড়কের ওপর নেতা-কর্মীদের বসার জন্য প্রায় ১ হাজার ২০০ চেয়ার বসানো হয়।

ঢাকা-চট্টগ্রাম মহাসড়কের ওপর মঞ্চ বানিয়ে গতকাল বুধবার সকালে সভা করেছে নারায়ণগঞ্জের সোনারগাঁ উপজেলা যুবলীগ এর ফলে গুরুত্বপূর্ণ মহাসড়কটিতে কিলোমিটারজুড়ে যানবাহনের সারি তৈরি হয় গতকাল সকাল মঞ্চের সামনে মহাসড়কের ওপর নেতা-কর্মীদের বসার জন্য প্রায় হাজার চেয়ার বসানো হয়।

‘ঢাকা’, ’চট্টগ্রাম’, ’ মহাসড়কের’, ‘ওপর’, ’ মঞ্চ’, ’ বানিয়ে’, ’ গতকাল’, ’ বুধবার’, ’ সকালে’, ’ সভা’, ’ করেছে’, ’ নারায়ণগঞ্জের’, ’ সোনারগাঁ’, ’ উপজেলা’, ’ যুবলীগ’, ’ এর’, ’ ফলে’, ’ গুরুত্বপূর্ণ’, ’ মহাসড়কটিতে’, ’কিলোমিটারজুড়ে’, ’ যানবাহনের’, ’সারি’, ’ তৈরি’, ’ হয়’, ’ গতকাল’, ’ সকাল’, ’ সাড়ে’, ’ সাতটায়’, ’ মঞ্চের’, ’ সামনে’, ’ মহাসড়কের’, ’ ওপর’, ’ নেতা’, ’কর্মীদের’, ’ বসার’, ’ জন্য’, ’ হাজার ‘, ’চেয়ার’, ’ বসানো’, ’ হয়’

Figure 3.3 Test preprocessing of a phrase.

3.5 Vector Encoding

By word Embedding, vector encoding of the words is achieved. The deep learning algorithms of Natural Language Processing are unable to interpret textual results. Word embedding here has a tremendous role to play in making it sensible for the neural network by uniquely specifying an integer vector for each word. For each vocabulary, it is essentially a 2D vector where vocabulary means rows and columns represent corresponding integers. In textual analysis, the larger embedding size illustrates more detailed relationships between terms and yields better performance. To date, any rich pre-trained Bengali language embedding is not available on the web. So we've used our own word embedding with 256 embedding dimensions of around 7500 vocabulary. A snip of pre-trained word embedding is outlined in table 3.3 on the next page.

Table 3.3 A snip of pre-trained embedding of words.

সময়	0.0713	-1.6886	2.0468	-1.023	-0.616	-2.621	1.62177	0.1077	-1.46	-1.030
থাকে	-1.4230	1.41716	1.9259	1.00332	-1.185	-1.2140	1.35827	0.48385	-1.300	0.57008
পর্যন্ত	-1.8481	0.1628	-1.2672	-0.8246	0.2001	1.1979	0.61423	-0.0566	-0.551	0.92260
আবারও	-0.6051	-0.5536	0.65958	0.8611	-1.332	0.04647	-0.0566	-0.5514	1.8330	0.64281
একজন	-0.2871	-0.2179	0.1711	-1.4828	-0.422	0.53509	0.47058	0.26869	-1.483	1.46605
করবে	2.14932	-0.7410	0.13895	1.93817	-1.483	-0.7888	1.46605	-1.4693	-0.592	-1.5522

3.6 Model settings

Based on the maximum sentence duration, we added padding to the phrase to balance our results. This is a length we set as 60 words. Padding makes the words equal in duration. In the LSTM sheet, we set 128 neurons and drop out value as 0,5 to minimize overfitting.

3.7 Train, Validation & Test

We used 90 percent of the 200 preprocessed documents containing around 2300 sentences to train our model. 5% is used to verify the outcome during preparation. After multiple epochs, the total value precision was 67 percent. The remaining 5% used

our model success to try it. After completing the training, model gives the probability of every sentence of given document. From that probabilities the summary will be generated in next process.

3.8 Summary generation

In fact, the model's given result is the percentage of probability for each sentence that guarantees its existence in the description. By preserving the chronological order of the original texts, we selected 5 most possible sentences and summarized them. The length of the description is based on consumer choices. Table 3.4 gives an example of sentences with expected probabilities.

Table 3.4 With expected odds, some words.

Sentences	Probabilities
তিনি বিষয়টি তদন্ত করে জানতে পারেন এবং দারিদ্রের কারণে দ্রুত মেয়ের বিয়ে দিতে চান তিনি	0.853
তবে ওই মেয়েকে স্বাবলম্বী হতে সহযোগিতা করেছে প্রশাসন	0.399
স্থানীয় প্রশাসনের পক্ষ থেকে বাল্যবিবাহ প্রতিরোধে এমন আরও নানা উদ্যোগ নেওয়া হয়েছে	0.602
জেলার কোথাও বাল্যবিবাহের সংবাদ পাওয়া গেলে তাৎক্ষণিকভাবে সেখানে যাচ্ছেন প্রশাসনের কর্মকর্তারা	0.745
গত ছয় মাসে প্রশাসনের হস্তক্ষেপে জেলায় ২০০টির বেশি বাল্যবিবাহ বন্ধ হয়েছে	0.561
প্রশাসন সূত্র জানায়, বাল্যবিবাহ কমাতে জেলা প্রশাসক মো. শফিকুল ইসলামের নেতৃত্বে কর্মপরিকল্পনা তৈরি করা হয়েছে	0.717

3.9 Additional model

For comparative analysis by GRU-RNN, where GRU stands for Gated Recurrent Unit[21], we have developed an additional model. This is also a modified version of the Recurrent Neural Network. There are two gates called upgrade gate & rest receive, where there are three gates for LSTM. It also offered optimal efficiency, but less than our suggested model. The effect of this model will be compared to our model in the next chapter, in the comparative analysis section.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results

We used 10 news article documents for experimenting and analyzing our model. Every document has a human generated summary. To evaluate our model, Rouge score values of Rouge-1, Rouge-2 and Rouge-3 were used, that is actually calculated from Recall, Precision & F-1. The experimental result of the model is described in Table 4.1.

Table 4.1 Output of our model in numerous Rouge ratings

Documents no.	Rouge-I			Rouge-II			Rouge-III		
	Recall	Precision	F-1	Recall	Precision	F-1	Recall	Precision	F-1
Doc-1	0.73	0.48	0.58	0.66	0.43	0.52	0.64	0.41	0.50
Doc-2	0.73	0.69	0.71	0.69	0.65	0.67	0.67	0.63	0.65
Doc-3	0.82	0.81	0.82	0.78	0.77	0.77	0.76	0.74	0.75
Doc-4	0.53	0.58	0.55	0.51	0.56	0.53	0.49	0.53	0.51
Doc-5	0.47	0.38	0.42	0.38	0.30	0.33	0.37	0.29	0.33
Doc-6	0.55	0.61	0.58	0.511	0.57	0.53	0.48	0.53	0.50
Doc-7	0.5	0.45	0.47	0.42	0.39	0.40	0.40	0.36	0.38
Doc-8	0.69	0.70	0.69	0.67	0.69	0.68	0.65	0.67	0.66
Doc-9	0.85	0.76	**0.80**	0.82	0.73	**0.78**	0.80	0.72	**0.76**
Doc-10	0.69	0.66	0.67	0.62	0.59	0.61	0.58	0.56	0.57

Table 4.1 The recall, consistency and F-1 score of such documents in various Rouge measures are demonstrated. Where we have the best F-1 ratings- 0.80, 0.78 & 0.76 consecutively for Rouge-I, Rouge-II and Rouge-III. Table 4.2 indicates the average scores in the review of those papers.

Table 4.2 Different Rouge average scores

Rouge	Recall (avg.)	Precision (avg.)	F-1 (avg.)
Rouge-1	0.666	0.60	0.633
Rouge-2	0.610	0.57	0.60
Rouge-3	0.590	0.53	0.57

4.2 Qualitative Analysis

We used very limited data sizes (only 200 documents), but any deep learning model requires a large amount of data for the best results. More preparation yields results that are more detailed. In this case, while our model has been trained with a relatively lower amount of data, it provides a surprising performance level. In order to preserve the length ration of model produced and referenced summaries, In summary, we selected documents that set five lengths of sentences for test selection. Table 4.3 displays a human-produced summary document and our model-generated summary, where the summary generated by our model is very similar to the reference summary.

Table 4.3 Example description created by our model

Document: মেহেরপুরের গাংনী উপজেলায় এক মায়ের চার মেয়ে। গত ২৪ জানুয়ারি স্কুলপড়ুয়া তৃতীয় মেয়ের বিয়ে দেওয়ার অনুমতি চেয়ে জেলা প্রশাসক বরাবর আবেদন করেন ওই মা। তাতে বলা হয়,স্কুল সনদ অনুযায়ী মেয়ের জন্মসাল ২০০৪। কিন্তু প্রকৃতপক্ষে মেয়ের বয়স ১৯ বছর।জেলা প্রশাসক মো. শফিকুল ইসলাম বিয়ের অনুমতি দেননি। তিনি বিষয়টি তদন্ত করে জানতে পারেন,দারিদ্রের কারণে দ্রুত মেয়ের বিয়ে দিতে চান তিনি। তবে ওই মেয়েকে স্বাবলম্বী হতে সহযোগিতা করেছে প্রশাসন। এটি বাল্যবিবাহ প্রতিরোধের একটি ঘটনা। স্থানীয় প্রশাসনের পক্ষ থেকে বাল্যবিবাহ প্রতিরোধে এমন আরও নানা উদ্যোগ নেওয়া হয়েছে । জেলার কোথাও বাল্যবিবাহের সংবাদ পাওয়া গেলে তাৎক্ষণিকভাবে সেখানে যাচ্ছেন প্রশাসনের কর্মকর্তারা। গত ছয় মাসে প্রশাসনের হস্তক্ষেপে জেলায় ২০০টির বেশি বাল্যবিবাহ বন্ধ হয়েছে। পাশাপাশি ২৯ জনকে বিভিন্ন মেয়াদে কারাদণ্ড দেওয়া হয়েছে। তাঁদের মধ্যে ঘটক,কাজি,বর ও অভিভাবকেরা আছেন। এ ছাড়া ৫৮ জনকে ৫৫ হাজার টাকা জরিমানা করা হয়েছে। মেহেরপুরে বাল্যবিবাহের হার তুলনামূলক বেশি। প্রশাসন সূত্র জানায়,বাল্যবিবাহ কমাতে জেলা প্রশাসক মো. শফিকুল ইসলামের নেতৃত্বে কর্মপরিকল্পনা তৈরি করা হয়েছে। পরিকল্পনার অংশ হিসেবে জেলা, উপজেলা,ইউনিয়ন ও ওয়ার্ড পর্যায়ে বাল্যবিবাহ প্রতিরোধ কমিটি গঠন করা হয়েছে। কমিটিতে নানা শ্রেণি-পেশার মানুষ রাখা হয়েছে। এসব কমিটি ছয় মাসে দুই শতাধিক মতবিনিময় সভা করেছে। ১৮টি ইউনিয়ন ও ১৬২টি ওয়ার্ডে সমাবেশের মাধ্যমে সাধারণ মানুষ ও প্রতিষ্ঠানের ব্যক্তিদের বাল্যবিবাহ প্রতিরোধে শপথ করানো হয়েছে।
Model generated: মেহেরপুরের গাংনী উপজেলায় এক মায়ের চার মেয়ে গত ২৪ জানুয়ারি স্কুলপড়ুয়া তৃতীয় মেয়ের বিয়ে দেওয়ার অনুমতি চেয়ে জেলা প্রশাসক বরাবর আবেদন করেন ওই মা তিনি বিষয়টি তদন্ত করে জানতে পারেন,দারিদ্রের কারণে দ্রুত মেয়ের বিয়ে দিতে চান তিনি জেলার কোথাও বাল্যবিবাহের সংবাদ পাওয়া গেলে তাৎক্ষণিকভাবে সেখানে যাচ্ছেন প্রশাসনের কর্মকর্তারা প্রশাসন সূত্র জানায়,বাল্যবিবাহ কমাতে জেলা প্রশাসক মো. শফিকুল ইসলামের নেতৃত্বে কর্মপরিকল্পনা তৈরি করা হয়েছে পরিকল্পনার অংশ হিসেবে জেলা,উপজেলা,ইউনিয়ন ও ওয়ার্ড পর্যায়ে বাল্যবিবাহ প্রতিরোধ কমিটি গঠন করা হয়েছে ১৮টি ইউনিয়ন ও ১৬২টি ওয়ার্ডে সমাবেশের মাধ্যমে সাধারণ মানুষ ও প্রতিষ্ঠানের ব্যক্তিদের বাল্যবিবাহ প্রতিরোধে শপথ করানো হয়েছে
Human generated: মেহেরপুরের গাংনী উপজেলায় গত ২৪ জানুয়ারি স্কুলপড়ুয়া তৃতীয় মেয়ের বিয়ে দেওয়ার অনুমতি চেয়ে জেলা প্রশাসক বরাবর আবেদন করেন ওই মা তিনি বিষয়টি তদন্ত করে জানতে পারেন,দারিদ্রের কারণে দ্রুত মেয়ের বিয়ে দিতে চান তিনি তবে ওই মেয়েকে স্বাবলম্বী হতে সহযোগিতা করেছে প্রশাসন স্থানীয় প্রশাসনের পক্ষ থেকে বাল্যবিবাহ প্রতিরোধে এমন আরও নানা উদ্যোগ নেওয়া হয়েছে জেলার কোথাও বাল্যবিবাহের সংবাদ পাওয়া গেলে তাৎক্ষণিকভাবে সেখানে যাচ্ছেন প্রশাসনের কর্মকর্তারা গত ছয় মাসে প্রশাসনের হস্তক্ষেপে জেলায় ২০০টির বেশি বাল্যবিবাহ বন্ধ হয়েছে প্রশাসন সূত্র জানায়,বাল্যবিবাহ কমাতে জেলা প্রশাসক মো. শফিকুল ইসলামের নেতৃত্বে কর্মপরিকল্পনা তৈরি করা হয়েছে

Since our model is supervised learning based, its performance is highly

dependent on the accuracy of training data. If the summaries of the dataset are prepared by the experts, the model will predict close to their cogitation. By overcoming data size and quality limitations, it achieved a very satisfied results.

Comparative Analysis

In the previous section, the outcome of our proposed model has already been discussed, now we want to make a comparison first of all between our model (LSTM) and the GRU-RNN model we originally developed. For Rouge-I, Rouge-II and Rouge-III, the average F-1 score was compared using the bar diagram in Figure 4.1 below:

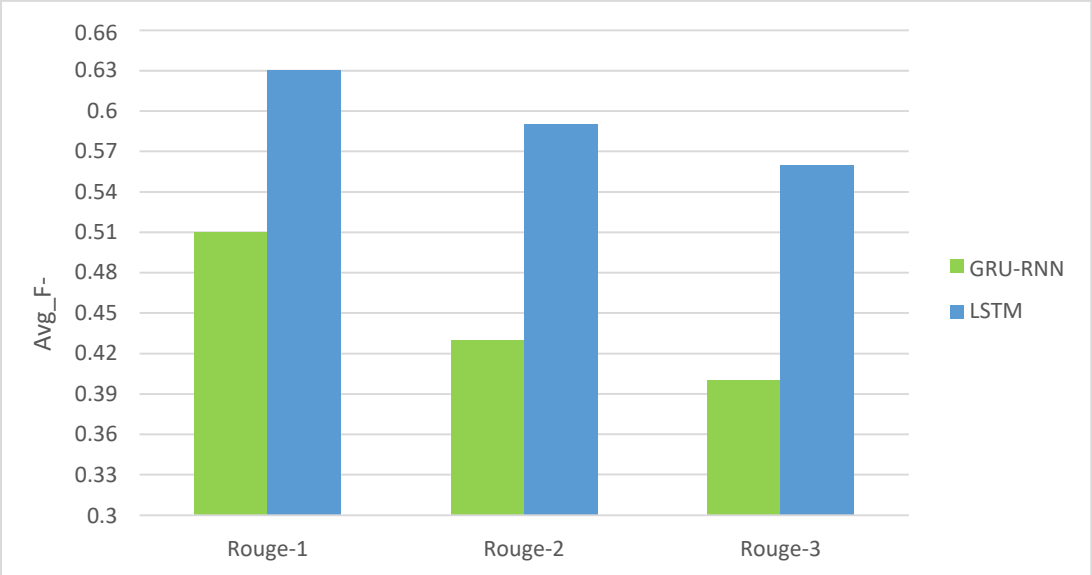


Figure 4.1 Comparison between the GRU-RNN outcome and LSTM

Figure 4.1 That means our model's efficiency (using LSTM) is much higher than the model based on GRU-RNN. In Rouge-I, the LSTM score is 24 percent higher than GRU-RNN, with Roue-II and Rouge-III respectively rising by 37 percent and 40 percent.

Second, we compared the outcome with two current models of Bengali text summarization[22,23] adopted in previous years. In order to validate these approaches and also our model, the same dataset[17] was used. In Figure-4.2, a comparative outcome of the average F-1 measure based on Rouge-1 and Rouge-2 was delineated. Models 1 in [22] and 2 in [23] are denoted.

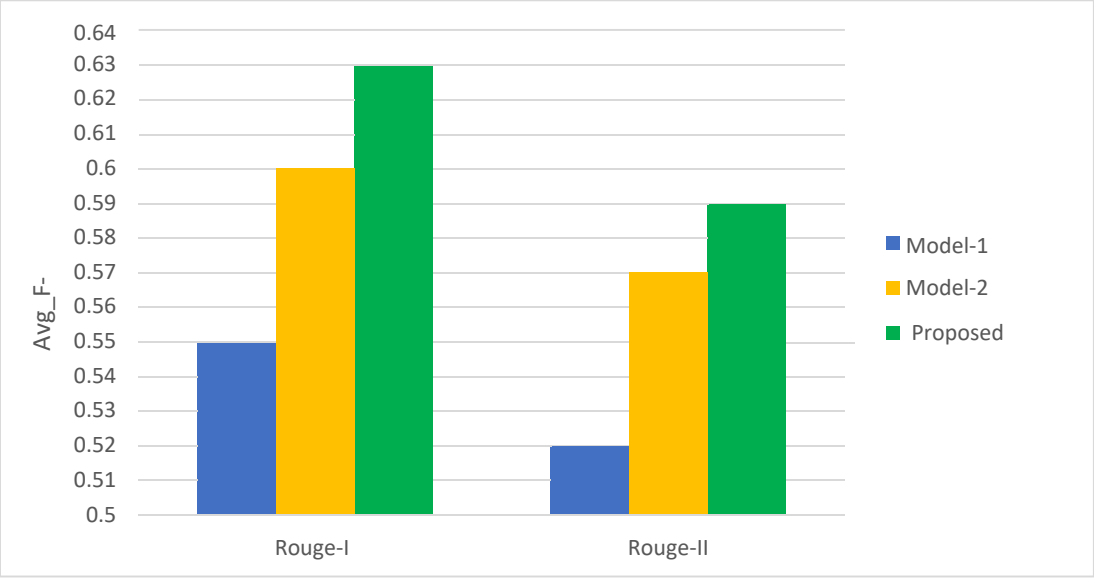


Figure 4.2 Comparative study of the average F-1 performance for the three Rouge-I and Rouge-II models.

The above comparison (in Figure 4.2) gives a clear understanding of our proposed model's performance. Based on the Rouge-1 & Rouge-2, the average F-1 scores are 3-6 percent higher than the nearest maximum score. We also measure Rouge-3 to display the precision more accurately. The average F-1 score of our Rouge-3 is 0.57, which is very good for the Bengali summary of the single document so far.

CHAPTER 5
SUMMARY, CONCLUSION, RECOMMENDATION & IMPLICATION FOR FUTURE RESEARCH

5.1 Summary

In either analysis, for the Bengali single paper, we proposed a straight forward deep neural network of sequence classification-based phrase extractive summarizer model. For English and other languages, there are many research articles on automated summary processes, but very few for the Bengali languageIn terms of our analysis, most of the Bengali summaries are rule-based and context-specific. Bengali is a complex language compared to English, due to its grammatical configuration, complicated alphabet, sentence formation and more. It is also usually very difficult to set the conventional description algorithm.

5.2 Conclusion

We implemented a deep neural network with a Bengali text summarizer in this case. This approach to this area is very recent. Where thousands to millions of data sets are required for supervised testing, we have very limited data. In spite of those inadequacy we have achieved satisfactory performance.

5.3 Recommendation

If this model trains with a large amount of quality data set, it will certainly improve precision and efficiency. There can be used properly annotated dataset to increase accuracy. The forecast for the paper under the particular domain would be more accurate for the domain-specific train collection.

5.4 Implication for Future Research

We used only one LSTM layer, but the result could be further enhanced by increasing the number of LSTM layers. There is an alternate alternative to maximize the outcome by using larger pre-trained word embedding sizes. Another scope is optimized the model for multi documents summarization.

REFERENCES

[1] Hinge, Sonam, and Sheetal Sonawane. "Cluster Based And Graph Based Methods Of Summarization: Survey And Approach." *International Journal of Computer Engineering and Applications* 10.II: 25-34.

[2] Al Munzir, Abdullah, et al. "Text analysis for Bengali Text Summarization using Deep Learning." *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019.

[3] Hahn, Udo, and Inderjeet Mani. "The challenges of automatic summarization." *Computer* 33.11 (2000): 29-36.

[4] Verma, Sukriti, and Vagisha Nidhi. "Extractive Summarization using Deep Learning." *arXiv preprint arXiv:1708.04439* (2017).

[5] Sarkar, Kamal. "Bengali text summarization by sentence extraction." *arXiv preprint arXiv:1201.2240* (2012).

[6] Abujar, Sheikh, et al. "A heuristic approach of text summarization for Bengali documentation." *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2017.

[7] Lin, C. Y., and E. Hovy. "Automated Text Summarization and the SUMMARIST System." *Proceedings of the TIPSTER Text Program* (1998): 197-214.

[8] Fattah, Mohamed Abdel, and Fuji Ren. "Automatic text summarization." *World Academy of Science, Engineering and Technology* 37 (2008): 2008.

[9] Akter, Sumya, et al. "An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm." *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2017.

[10] Bangla Newspaper List of all Online Bangladeshi Newspaper https://www.24livenewspaper.com/bangla-newspaper Accessed on 29 march 2019.

[11] Luhn, Hans Peter. "The automatic creation of literature abstracts." *IBM Journal of research and development* 2.2 (1958): 159-165.

[12] Baxendale, Phyllis B. "Machine-made index for technical literature—an experiment." *IBM Journal of research and development* 2.4 (1958): 354-361.

[13] Cheng, Jianpeng, and Mirella Lapata. "Neural summarization by extracting sentences and words." *arXiv preprint arXiv:1603.07252* (2016).

[14] Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou. "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents." *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[15] Cao, Ziqiang, et al. "Ranking with recursive neural networks and its application to multi-document summarization." *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[16] Isonuma, Masaru, et al. "Extractive summarization using multi-task learning with document classification." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

[17] Das, Amitava, and Sivaji Bandyopadhyay. "Topic-based Bengali opinion summarization." *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010.

[18] Bangla Natural Language Processing Community, http://bnlpc.org/research.php, accessed on 01 march, 2019

[19] jellyfish 0.7.1 python library. https://pypi.org/project/jellyfish, Accessed on 20 march, 2019

[20] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[21] Bengali Stop words list. https://www.ranks.nl/stopwords/bengali, Accessed on 10 march, 2019

[22] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[23] K. Sarkar, "A keyphrase-based approach to text summarization for English and Bengali documents," *International Journal of Technology Diffusion (IJTD)*, vol. 5, no. 2, pp. 28-38, 2014.

[24] Haque, Md Majharul, Suraiya Pervin, and Zerina Begum. "Automatic Bengali news documents summarization by introducing sentence frequency and clustering." *2015 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE,2015.

Text analysis for Bengali long text summarization using deep learning

ORIGINALITY REPORT

17%	8%	10%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Abdullah Al Munzir, Md. Lutfor Rahman, Sheikh Abujar, Ohidujjaman, Syed Akhter Hossain. "Text analysis for Bengali Text Summarization using Deep Learning", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019 Publication	8%
2	Submitted to Daffodil International University Student Paper	7%
3	Submitted to Malaviya National Institute of Technology Student Paper	1%
4	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
5	aclweb.org Internet Source	<1%
6	dspace.jaist.ac.jp Internet Source	<1%