INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

IJARnD

(Volume 4, Issue 11)
Available online at: www.ijarnd.com

Marathi text summarization using neural networks

Anishka Chaudhari[1], Akash Dole[2], Deepali Kadam[3]
Student[1,2], Professor[3], Datta Meghe College of Engineering, Navi Mumbai, Maharashtra

ABSTRACT

The internet is comprised of web pages, news articles, status updates, blogs and much more. It is difficult to navigate through this data as it is unstructured and usually discursive. Condensed versions of this data are generated so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. We propose a system for extractive text summarization method using neural networks for Marathi text. Extractive summaries or extracts are produced by identifying important sentences or words which are directly selected from the document. To perform extractive text summarization we propose to use a Recurrent Neural Network (RNN) – a type of neural network that can perform calculations on sequential data (e.g. sequences of words) – as it has become the standard approach for many Natural Language Processing tasks. The translation of the Marathi text to English will be done using the Google translate API for this proposed system.

Keywords— *Text summarizer, Google translate, Neural machine translation, NLP, Neural Networks, Encoder-Decoder, Bahdanau Attention model, Extractive summarization*

1. INTRODUCTION

The main use of text summarization is to help readers save time and effort of reading long documents to find useful information. Text summarization is creating a short, accurate and consistent summary of a longer document. We cannot create summaries of all the text manually and there is a great demand for automatic method. Summarization helps users in several ways such as reducing reading time, making selection procedures faster and easier for researching documents, improves effectiveness of indexing. Automatic summaries are less biased as compared to human summaries. Summarization programs and systems are commercially in demand by the military, universities, research institutes, law firms, etc.

2. CONCEPT OF TEXT SUMMARIZATION

There are mainly 3 types of summarizations – Abstraction based, Extraction based and Aided summarization.

2.1 Abstraction based summarization

Abstraction based summaries are similar to human-level summaries. This method involves paraphrasing sections of the source document. Abstraction can condense the text more strongly than any other method but the system that does this is very sophisticated and difficult to develop as the technology and field are ever-growing.

2.2 Extraction based summarization

Extractive summarization is when the system or program selects sentences or words directly from the original document depending on their importance without modifying the sentences. The Marathi text summarization system that we are proposing is based on this method.

2.3 Aided summarization

Machine Learning techniques that have been adapted from fields such as text mining or information retrieval are used for aided summarization.

3. PROBLEM STATEMENT

As the amount of information available on the web is getting double day by day which is leading to information overload. Finding important and useful information is becoming difficult. The automatic summary generation technique addresses the issue of generating shortened information from documents written on the same topic. It offers a possibility of finding main points of texts and so the user can save time on reading whole document.

4. OBJECTIVE AND PURPOSE

The objective of the project to implement a text summarizer that summarizes Marathi news articles using neural networks. As the amount of information available on the web is getting double day by day which is leading to information overload. Finding important and useful information is becoming difficult. The automatic summary generation technique addresses the issue of generating shortened information from documents written on the same topic. It offers a possibility of finding main points of texts and so the user can save time on reading whole document.

5. PROPOSED SYSTEM AND METHOD

The main idea of the proposed system is that we translate the input Marathi text or documents of the user to English. The translated text in English is then sent to the summarizer. Then the summarized text is again translated back to Marathi language and given to the user as a result. The accuracy of the translation is to be measured using the BLEU score and the accuracy of the summary is measured using Recall. The proposed system is explained in detail further.

5.1 Overview

We first translate our Marathi dataset to English using Google Translate API and then summarize news articles using a bi-directional encoder-decoder LSTM model. The resultant summary is again translated to Marathi using Google Translate API.

5.2 Translation phase

We used Google Translate to translate our dataset from Marathi to English. We uploaded pairs of Marathi news articles and their summaries to be translated into English for the translation phase.

5.3 Understanding the Encoder-Decoder Architecture

Let's understand this from the perspective of text summarization. The input is a long sequence of words and the output will be a short version of the input sequence.

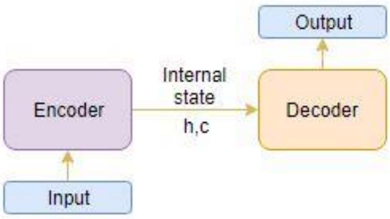

Fig. 1: Encoder-Decoder Architecture

Generally, variants of Recurrent Neural Networks (RNNs), i.e. Gated Recurrent Neural Network (GRU) or Long Short Term Memory (LSTM), are preferred as the encoder and decoder components. This is because they are capable of capturing long term dependencies by overcoming the problem of vanishing gradient. We can set up the Encoder-Decoder in 2 phases:

- Training phase
- Inference phase

5.4 Training phase

In this stage, we will initially set up the encoder and decoder. We will at that point train the model to foresee the objective grouping balanced by one timestep.

5.4.1 Encoder: An Encoder Long Short Term Memory model (LSTM) reads the entire input sequence then, at each timestep, one word is fed into the encoder. It then processes the information at every timestep and captures the contextual information present in the input sequence. We have put together the below diagram which illustrates this process:

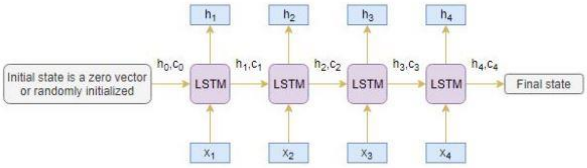

Fig. 2: Encoder

The hidden state (hi) and cell state (ci) of the last time step are used to initialize the decoder. This is because the encoder and decoder are two different sets of LSTM architecture.

5.4.2 Decoder: The decoder is also an LSTM network that reads the entire target sequence word-by-word and predicts the same sequence offset by one timestep. The decoder is trained to predict the next word in the sequence given the previous word.

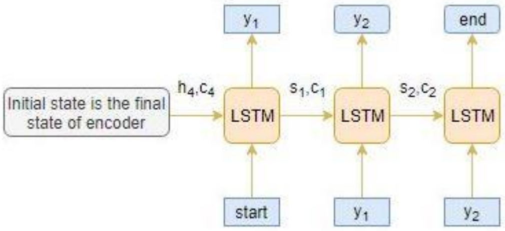

Fig. 3: Decoder

<start> and <end> are the special tokens that are added to the target sequence before feeding it into the decoder. The target sequence is unknown while decoding the test sequence. So, we start predicting the target sequence bypassing the first word into the decoder which would be always the <start> token. And the <end> token signals the end of the sentence.

5.5 Inference Phase

After training, the model is tested on new source sequences for which the target sequence is unknown. So, we need to set up the inference architecture to decode a test sequence:

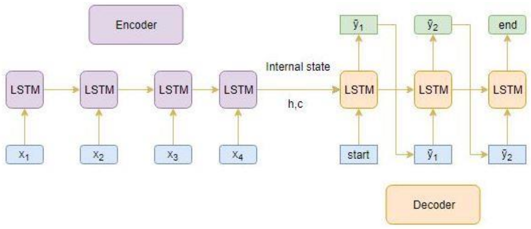

Fig. 4: Inference Phase

5.6 Attention Mechanism

The proposed system implements Bahdanau-style (additive) attention. We used the BahdanauAttention class from Tensorflow (Seq2Seq).

6. EVALUATION RESULTS

The dataset contained 1000 Marathi news articles with corresponding summaries. The proposed method achieved mean average precision of 0.319 and overall recall of 0.342. The overall accuracy was 0.323. The precision could be probably improved by using a different model. Although the results are not very good this does not rule out the possibility of using deep learning for the task.

7. CONCLUSION

Summaries can be used to implement a document retrieval system as they can be used as a representation of the document needed. This paper presents an extraction system using LSTM network. The results are not great but LSTM is suited to the task at hand and should be explored further.

8. REFERENCES

[1] Ms. Deepali Kadam, "International Journal of Innovations and; Advancement in Computer Science" IJIACS ISSN 2347 – 8616 Volume 4, Special Issue March 2015.

[2] Ramesh Nallapati, Feifei Zhai, Bowen Zhou. 2016. SummaRuNNer: A Recurrent Neural Network-based Sequence Model for Extractive Summarization of Documents arXiv: 1611.04230v1 [cs.CL] 14 Nov 2016

[3] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. ArXiv: 1602.06023v5 [cs.CL], 2016.

[4] D. K. Kanitha, D. Muhammad Noorul Mubarak &amp; S. A. Shanavas. 2018. COMPARISON OF TEXT SUMMARIZER IN INDIAN LANGUAGES "International Journal of Advanced Trends in Engineering and Technology (IJATET)" Volume 3, Issue 1

[5] Virat V. Giri, Dr.M.M. Math, and Dr.U.P. Kulkarni, A Survey of Automatic Text Summarization System for Different Regional Language in India, Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6, Special Issue, October 2016.

[6] Aliaksei Severyn and Alessandro Moschitti, Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks, http://citeseerx.ist.psu.edu

[7] Tarasov D. S., Natural Language Generation, Paraphrasing and Summarization of User Reviews with Recurrent Neural Networks, ReviewDot Research, Kazan, Russia.

[8] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, Chandan K. Reddy, Senior Member, IEEE, Neural Abstractive Text Summarization with Sequence-to-Sequence Models, arXiv:1812.02303v2 [cs.CL] 7 Dec 2018.

[9] Rupal Bhargava, Sukrut Nigwekar, Yashvardhan Sharma, Catchphrase Extraction from Legal Documents Using LSTM Networks, http://ceur-ws.org/Vol-2036/