Extractive Text Summarization with Teacher Forcing Algorithm

Shimul Paul
160204014@aust.edu

Amit Saha Ami
160204016@aust.edu

Zakia Sultana Tanzil
160204004@aust.edu

1. Introduction

Preserving the key elements and the actual meaning of a content but reducing the texts to the shorter version is called summarization of that content.There are two types of summarization a)Extractive Summarization b)Abstractive Summarization.

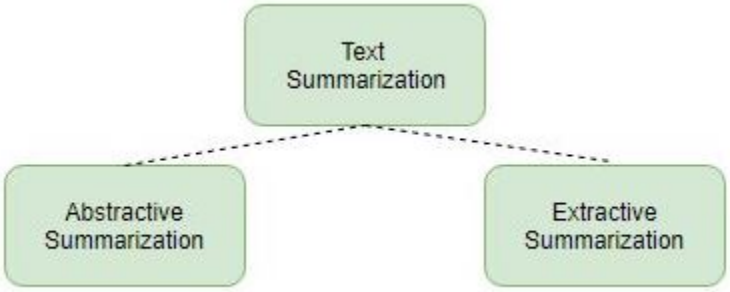

Fig 1 : Two Types of Summarization

Extractive Summarization: A subset of words that represent the most important points pulling from a piece of text and combining to make a summary is called an **Extractive Summarization.**

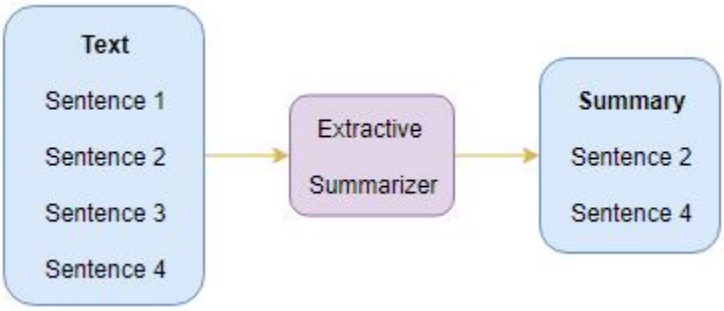

Fig 2 : Method of Extractive Summarization

Abstractive Summarization: It is the procedure of generating a summary of a text from its main ideas, not by copying verbatim most salient sentences from text. This is an important and challenging task in natural language processing.

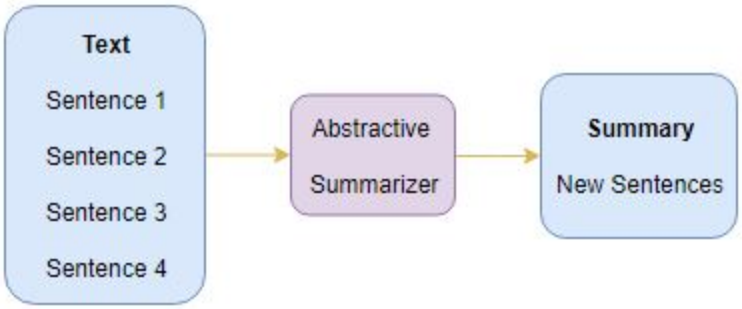

Fig 3 : Method of Abstractive Summarization

An example of summary of a content is given below in extractive based summarization:

Source Text:

Peter and Elizabeth took a taxi to **attend** the night **party** in the **city**.While in the party,**Elizabeth** collapsed and was **rushed** to the **hospital**.

Summary:

Peter and Elizabeth attend party city.Elizabeth rushed hospital.

We tried implementing **extractive** summarization which selects passages from the source text and then arranges it to form a summary.

We are trying to solve this as summarizing a text manually is time costly and strenuous. The computerization of this task is being loved magnificently in this big data era where there has been voluminous text data from many sources. The data is an incredible source of knowledge but in this busy world people don't have that much time to go through the entire content to extract the useful or important portion. So, this is important to summarize the voluminous data automatically in an effective manner.

There are various important applications relying on text summarization such as question answering, news summarization, heading generation, legal text summarization etc.

c. The task is challenging, as when a content is summarized by us as humans, generally we go through it completely to build up our understanding, and then write a summary spotlighting its key points. As computers don't have earthling knowledge and dialect aptness, it assembles self-regulating text summarization, a hard and substantial task.

2. Related Works

List of some important related works :

Web summarization

Web pages carry lots of portions which cannot be summarized such as pictures. The textual details they have is often insufficient, which makes applying text summarization techniques bounded. The preliminary research in this regard is [1] where query web search engines and summon the pages having links to the designated web page. Then they examine the aspirant pages and choose the best sentences carrying links to the web page analytically. In paper [2] extended and refined this approach by using an algorithm trying to choose a sentence regarding the same topic that enfolds as many aspects of the web page as feasible. For blog summarization,in [3] proposed a procedure that first procures indicative words from comments and picks out important sentences from the blog post carrying indicative words.

Scientific Article Summarization

A beneficial source of information when summarizing a scientific paper (i.e. citation-based summarization) is to discover other papers that quote the target paper and pull out the sentences in which the recommendations take place for identifying the important aspects of the target paper. In paper [3] presents a language model that gives a likelihood to each word in the quotation context sentences. They then count the importance of sentences in the authentic paper using the KL divergence technique (i.e. finding the resemblance between a sentence and the language model).

Email Summarization

Email has some clear characteristics that specifies the aspects of both vocalized conversation and penned text. For example, summarization procedures must contemplate the communal nature of the chat as in vocalized conversations. In paper [4] proposed early research in this regard, by presenting a method to create a summary for the first two levels of the drift discussion. A thread comprises one or more conversations between two or more contributors over time. They choose a message from the source message from each response to the source, considering the intersection with source context. In paper [5] used a machine learning procedure and comprehended features related to the thread as well as properties of the email structure such as placement of the sentence in the tread, total of receivers, etc.

Woking Approach and Limitations:

1. **BeautifulSoup for text preprocessing**

   BeautifulSoup is a Python package for working with real-world and broken HTML, just like **lxml.html**. As of version 4.x, it can use **different HTML parsers**, each of which has its advantages and disadvantages .

   A very nice feature of BeautifulSoup is its excellent **support for encoding detection** which can provide better results for real-world HTML pages that do not (correctly) declare their encoding.

   Another good feature of **BeautifulSoup** is that It's very fast and lenient.

2. Some papers[3,5] used only encoder-decoder models for text summarization , but we have applied Attention Decoder RNN.

   Attention is a mechanism combined in the **RNN** allowing it to focus on certain parts of the input sequence when predicting a certain part of the output sequence, enabling easier learning and of higher quality.

   Attention is proposed as a solution to the limitation of the Encoder-**Decoder** model encoding the input sequence to one fixed length vector from which to **decode** each output time step. This issue is believed to be more of a problem when **decoding** long sequences.

3. **Teacher Forcing Algorithm**

   Teacher forcing is a method for quickly and efficiently training recurrent neural network models that use the ground truth from a prior time step as input.

   It is a network training method critical to the development of deep learning language models used in machine translation, text summarization, and image captioning, among many other applications.

   Let's make a teacher forcing concrete with a short worked example.

   Given the following input sequence:

   "mango is my favourite fruit"

   Imagine we want to train a model to generate the next word in the sequence given the previous sequence of words.

   First, we must add a token to signal the start of the sequence and another to signal the end of the sequence. We will use "[START]" and "[END]" respectively.

   [START] Mango is my favourite fruit [END]

Next, we feed the model "[START]" and let the model generate the next word.

Imagine the model generates the word "*is*", but of course, we expected "*mango*"

X,	yhat
[START],	is

Naively, we could feed in "*is*" as part of the input to generate the subsequent word in the sequence.

X,	yhat
[START], a,	?

You can see that the model is off track and is going to get punished for every subsequent word it generates. This makes learning slower and the model unstable.Instead, we can use teacher forcing.In the first example when the model generated "*is*" as output, we can discard this output after calculating error and feed in "*mango*" as part of the input on the subsequent time step.

X,	yhat
[START], Mango,	?

We can then repeat this process for each input-output pair of words.

X,	yhat
[START],	?
[START], mango,	?
[START], mango, is,	?
[START], mango, is, my,	?
...	

The model will learn the correct sequence, or correct statistical properties for the sequence, quickly.

3. Project Objective(s)

Tasks of our system

"The main objective of our system is to summarize an article"

Steps:

3.1. First we load the WikiHow dataset. This is in .csv format and has 2 columns(Headline, Article).
3.2. Text Preprocessing

Now we need to clean our text, we perform the following steps for the text and headlines pair:

- Remove extra white spaces
- Expand contractions
- Remove special characters
- Lowercase all texts

3.3. Before moving on to build our model, we need to do a test train split for our data. We use the sklearn to do the same. We will use 70 % of the data as training data and evaluate the performance on the remaining 30 %.

3.4. Encoder-Decoder Architecture

Let us see a high-level overview of Encoder-Decoder architecture and then see its detailed working in the training and inference phase.Intuitively this is what happens in our encoder-decoder network:

- We feed in our input (in our case text from news articles) to the Encoder unit. Encoder reads the input sequence and summarizes the information in something called the internal state vectors (in case of LSTM these are called the hidden state and cell state vectors).
- The encoder generates something called the context vector, which gets passed to the decoder unit as input. The outputs generated by the encoder are discarded and only the context vector is passed over to the decoder.
- The decoder unit generates an output sequence based on the context vector.

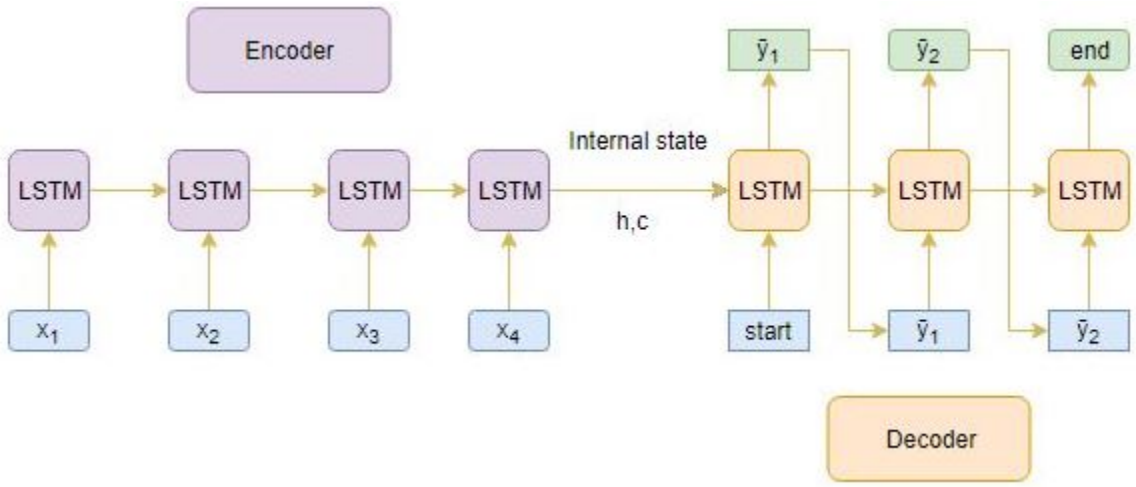

Fig 4: Encoder - Decoder Model for Text Summarization

3.5. Attention Decoder[8]:

- When humans read any lengthy paragraph, they pay attention to certain words then they change their attention to the next few words and so on.
- Intuitively think of your teacher correcting your 10 mark answer in your History paper. The teacher would have a key with them in which certain important points /words are there. So in your answer sheet your teacher would look for these important words in the key. More attention is paid to the keywords.
- Hence humans change their attention from one sequence of words to another sequence of words when given a lengthy input sequence
- So, instead of looking at all the words in the source sequence, we can increase the importance of specific parts of the source sequence that result in the target sequence. This is the basic idea behind the attention mechanism.
- Attention mechanism makes use of bidirectional RNNs The regular RNN is unidirectional as the sequence is processed from the first word to the last word. In bidirectional RNN we will have a connection in forward and reverse direction.
- So in addition to the forward connection, there is also a backward connection for each of the neurons.
- The outputs generated from the forward and the backward connection of neurons are concatenated together to give outputs y1^,y2^and so on. So we will have two back propagations one for the forward path in the backward direction and again for the backward path in the forward direction.
- The context vector is nothing but the weighted sum of outputs from the encoder.

There are 2 different classes of attention mechanism depending on the way the attended context vector is derived:

- Global Attention-Here, the attention is placed on all the source positions. In other words, **all the hidden states of the encoder are considered for deriving the attended context vector:**

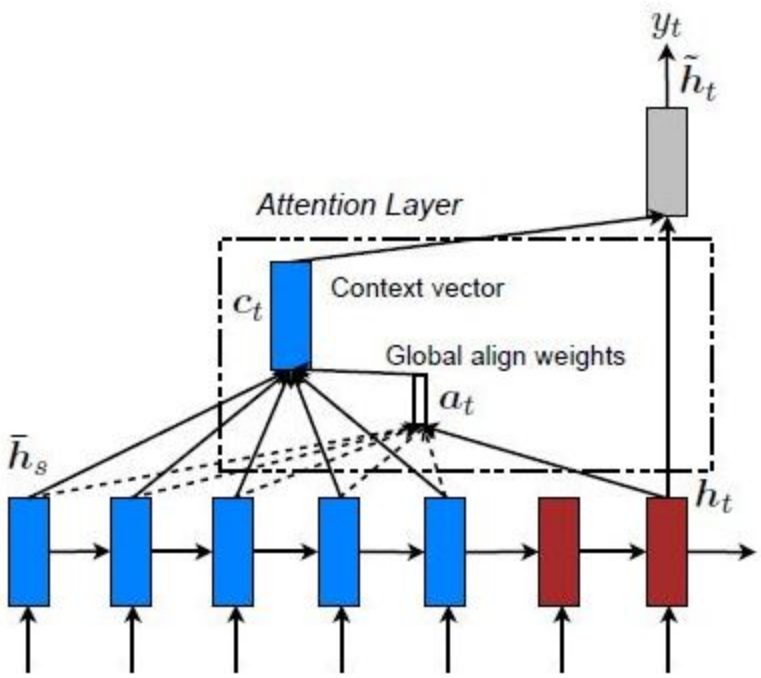

Fig 5: Structure of Global Attention

- Local Attention-Here, the attention is placed on only a few source positions. **Only a few hidden states of the encoder are considered for deriving the attended context vector.**

We will be using **global attention** for our task.

3.6.After that we extract the sentences from the article.Basically **Sentence extraction** is a technique used for automatic **summarization** of a text. In this shallow approach, statistical heuristics are used to identify the most salient **sentences** of a text.
3.7. Then our main part “Summarization”
3.8. For evaluating our system we applied the **Rouge Scoring** Method.

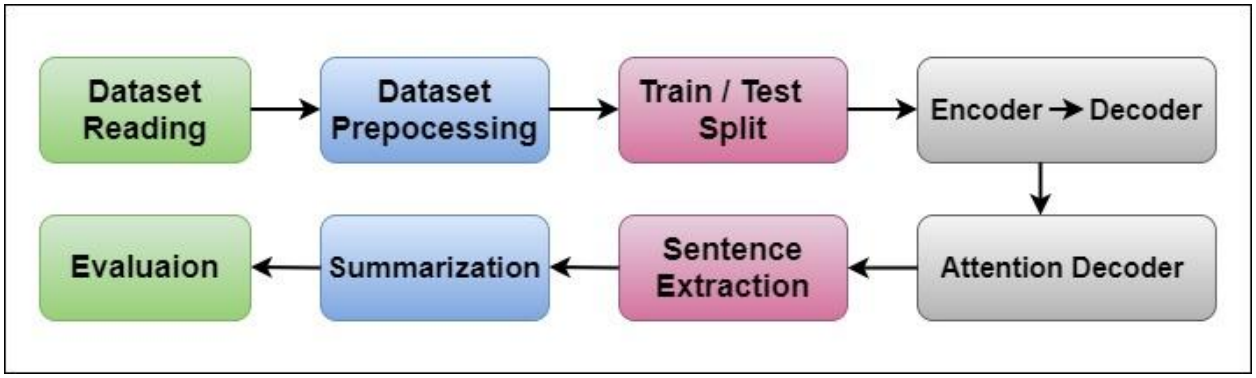

Fig 6:Flow Diagram of system

Table 1:Some dummy input and output of our system.

Sl No	Title	Article	Summary(Output)
1	sign free account book board click link email follow instructions complete profile finally create child profile	go http bookboard com enter email address receive confirmation email link complete account setup email arrive couple minutes check spam folder guide three easy steps first set password enter name avatar nickname choose nickname avatar specify child age birth month day gender fields optional want provide information enables book board personalize reading experience closely child age possible interests	open new store app create new account check email address new email name check email send new email <EOS>
2	check different races coming area time race properly pick race appeals register team members head race day read race regulations know time show report else need race day	communities races year long consider choosing spring fall first collect extreme temperatures slow cause additional discomfort run want minimum weeks train look races least two months away many races run charity organization cause want support sign race raises money mission pay required entrance fees usually	Collect money for charity organization want buy start get get kill usually <EOS>

3	change hairstyle use color play volume invest clip extensions	cannot control genetics use power illusion make hair appear thicker ask hairdresser best cut style create thickness volume hair add layers create fullness especially hair straight feeling daring consider getting bangs contrast bangs rest hair help accentuate length get trim hair tapers ends tends appear thin unhealthy trimming hair even length make hair appear healthier bouncy get rid split ends process keep hair looking salon ready getting trimmed every weeks highlights add extensions try clip ons much cheaper version hair extensions purchased local beauty supply stores find color texture similar clip daily basis add volume hair	find a similar color texture clip. make sure use hair green hair add add add hair <EOS>
4	access code page want redirect open code edits amend code save file upload old domain test redirect	different method involves changing webpage code directly first need download files associated url would like redirect away note cases using meta command ideal redirect web pages meta code redirects often filtered search engines since known spam technique use notepad similar text editor refresh content stands number seconds redirect happens www newsite com newurl html specific url page redirected also possible create custom error message announcement sites moved may draw unwanted attention redirect redirecting traffic old url likely changes also taken place url code important url code contains meta redirect code type url browser directly use search engine find page redirect immediately new url specified code without messages landing points	download install website launch music website launch music website download install music website <EOS>

4. Methodologies / Model

Steps to solve the problem and the diagrams of the models we used

We have discussed this part in section 3

5. Experiments

5.1. Dataset:

WikiHow: A Large Scale Text Summarization Dataset [12]

Each article consists of multiple paragraphs and each paragraph starts with a sentence summarizing it. By merging the paragraphs to form the article and the paragraph outlines to form the summary, the resulting version of the dataset contains more than 200,000 long-sequence pairs.

There are two separate data files containing the articles and their summaries:

The **wikihowAll.csv** file consisting of the concatenation of all paragraphs as the articles and the bold lines as the reference summaries :

Part	Description
Title	The title of the article as it appears on the WikiHow knowledge base
Headline	The concatenation of all the bold lines (the summary sentences) of all the paragraphs to serve as the reference summary
Text	The concatenation of all paragraphs (except the bold lines) to generate the article to be summarized

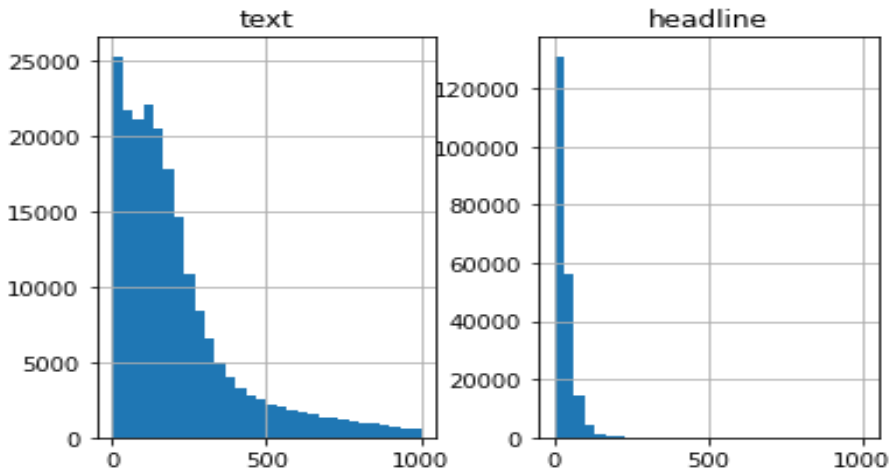

Fig 7: Word length frequency

Table 2: Samples from the dataset

Headline	Text
Keep related supplies in the same area.,Make an effort to clean a dedicated workspace after every session.,Place loose supplies in large, clearly visible containers.,Use clotheslines and clips to hang sketches, photos, and reference material.,Use every inch of the room for storage, especially vertical space.,Use chalkboard paint to make space for drafting ideas right on the walls.,Purchase a label maker to make your organization strategy semi-permanent., Make a habit of throwing out old, excess, or useless stuff each month.	If you're a photographer, keep all the necessary lens, cords, and batteries in the same quadrant of your home or studio. Paints should be kept with brushes, cleaner, and canvas, print supplies should be by the ink, etc. Make broader groups and areas for your supplies to make finding them easier, limiting your search to a much smaller area. Some ideas include: Essential supplies area -- the things you use every day. Inspiration and reference area. Dedicated work area . Infrequent or secondary supplies area, tucked out of the way.; , This doesn't mean cleaning the entire studio, it just means keeping the area immediately around the desk, easel, pottery wheel, etc. clean each night. Discard trash or unnecessary materials and wipe down dirty surfaces. Endeavor to leave the workspace in a way that you can sit down the next day and start working immediately, without having to do any work or tidying. Even if the rest of your studio is a bit disorganized, an organized workspace will help you get down to business every time you want to make art. , As visual people, a lot of artist clutter comes from a desire to keep track of supplies visually instead of tucked out of sight. By using jars, old glasses, vases, and cheap, clear plastic drawers, you can keep things in sight without leaving it strewn about haphazardly. Some ideas, beyond those just mentioned, include:

	Canvas shoe racks on the back of the door Wine racks with cups in each slot to hold pens/pencils. Plastic restaurant squirt bottles for paint, pigment, etc., Simply string up the wires across a wall or along the ceiling and use them to hold essential papers that you don't want to cut or ruin with tacks or tape. Cheap and easy, this is also a good way to handle papers and ideas you touch regularly or need to pin up and down for inspiration., Shelving is an artist's best friend and is a cheap and easy way to get more room in your studio or art space. Don't be afraid to get up high either, especially for infrequently used supplies. The upper reaches of the room are often the most under-utilized, but provide vital space for all your tools and materials., Turning one wall into a chalkboard gives you a perfect space for ideas, sketches, and planning without requiring extra equipment or space. You can even use it for smaller areas. Paint over jars or storage equipment, allowing you to relabel them with chalk as your needs change. , A lot of disorganization comes when you keep moving the location of things, trying to optimize your space by reorganizing frequently. This usually has the opposite effect, leading to lost items and uncertainty when cleaning, but an afternoon with a label maker can solve everything. Instead of spending all of your mental energy looking for or storing things, you can just follow the labels, freeing your mind to think about art., Once a month, do a purge of your studio. If it isn't essential or part of a project, either throw it out or file it away for later. Artists are constantly making new things, experimenting, and making a mess. This is a good thing, but only if you set aside time to declutter. It may not be fun at the moment, but it is a lot more fun than spending 30 minutes digging through junk to find the right paint or an old sketch. Don't be sentimental here. If you haven't used it in the last six months there is little chance you'll use it in the next six months. Toss it.
Make a list of what your friends watch, read, and listen to now., Go online to check celebrity gossip., Watch popular TV shows to be in the know., Check Top 100 charts to see what music is popular now., Go to the movies whenever something exciting is released., Follow your favorite celebrities on Twitter and Instagram., Vary your interests to avoid missing	Use your friendsâ€™ conversations to figure out whatâ€™s popular. Pay careful attention to which shows, bands and artists, movies, and celebrities they mention. Make a list and add new references whenever they come up. You can use this list to pick which people and sources of entertainment you should follow first. Itâ€™ll be most important for you to catch the references your friends drop, so this is a great place for you to begin your pop culture education.; , There are lots of websites and online magazines that detail the glitzy and glamorous lives of the rich and famous. Some teen magazines also have content on these topics. Use these to make sure you never miss a reference to

references., Listen to podcasts about pop culture.	someoneâ€™s new baby or a dramatic breakup.Download your favorite celebrity news outletâ€™s app to be instantly informed of new gossip. Try Us Weekly or People. Seventeen and Teen Vogue are also good sources. , TV shows are big sources of pop culture. Even if you donâ€™t have cable, you can stream shows through platforms like Hulu. Be sure to also check out Netflix for binge-able content your friends will surely talk about.If youâ€™re willing to wait a week for new episodes, you can usually watch popular shows either on the networksâ€™ websites or on Hulu for free. , Billboard.com and iTunes both have Top 100 charts that keep track of the most popular songs in the United States. Go online to look at these charts and download the top 10 or 20 songs. Youâ€™ll be singing along to all the hits in no time.Billboard updates their chart weekly, and iTunes updates in real time. , Youâ€™ll know when this happens. Even if people arenâ€™t planning Harry Potter-style midnight showing parties, make an event of big movie releases. Get your friends together on a Friday or Saturday night and do a group outing. , Social media has made it much easier to feel connected to the stars. Check Twitter to stay updated on their thoughts, opinions, and even their politics. Use Instagram to see their cute or funny photos.Check whatâ€™s trending on Twitter to see what people are talking about on a day-to-day basis. This is a great way to make sure you donâ€™t miss any viral stories that could be tomorrowâ€™s pop culture references. , As you start to follow new shows, celebrities, and artists, donâ€™t stick with one genre. You may decide you really like crime shows, but be sure to check out some dramas as well. This will help you avoid missing references that fall a little outside your normal set of interests., If youâ€™re a podcast listener, youâ€™re in luck! There are tons of shows dedicated to giving you weekly content just about pop culture. The hosts will make jokes about it, criticize it, and generally keep you informed.NPRâ€™s Pop Culture Happy Hour and Slateâ€™s Culture Gabfest are catch-all shows about all pop culture. Who Charted? discusses top music and movies. Nerd Culture Podcast focuses on all things nerd, including comics, TV shows, and movies.You can also follow pop culture writers on Twitter.
Read your local phone book., Determine the distance from your location to each casino., Choose the casino that is the shortest distance from your location.	Check for a section that is titled "Casino", or something similar. Note the location of each casino. , If you learn about more than 1 casino in your phone book, use the Internet to search the distance between your location and each casino. Sites such as

	maps.google.com or mapquest.com will help you in this search.
Look for a major or complete desertion of the colony by all adult bees over a short period of time., Look for bees., Check for the queen., Check for another colony moving in.	The period may be hours or days.; , If bees have been killed by something toxic or by something in the hive, they will still be present. , She will often still be present with a very reduced number of workers. , With CCD, this is unlikely to happen (robber bees). There may also be an unusual delay before wax moths invade the dead colony.

iii. Train / Test split.

We have divided our dataset into 2 segments

- Train: 70%
- Test: 30%

b. Evaluation Metric

i. Evaluation method for our proposed system

Rouge Score:ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially a set of metrics for evaluating automatic summarization of texts as well as machine translations.

It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced).

ii. List important evaluation metrics you choose to evaluate your system.

We have used only Rouge Score to evaluate our system

c. Results

i. Present the results with some tables, graphs etc and discuss results.

Table 3: Ablation experiment based on the hyperparameters.

Sl No	Train-Test ratio	RNN	Optimizer	Learning Rate	Hidden Node	Drop Out	Loss	Iterations
1	70-30	LSTM	Adam	0.001	200	0.2	3.56	40000
2	70-30	GRU	SGD	0.01	300	0.3	5.28	50000
3	70-30	GRU	SGD	0.001	300	0.1	5.47	50000
4	70-30	GRU	Adam	0.001	300	0.3	5.77	50000
5	70-30	GRU	Adam	0.001	200	0.25	5.83	50000
6	70-30	GRU	SGD	0.01	300	0.1	5.23	100000
7	70-30	LSTM	Adam	0.003	200	0.3	3.61	50000
8	70-30	LSTM	SGD	0.01	300	0.1	3.52	150000

Some rogue score :

- **Article**: choose weight ball challenging also move quickly quickly draw back return center alignment
  Headline: draw sure want draw
  Rouge Score: 0.11764705522491362
  Summary: draw sure want draw <EOS>

- **Article**: make sure every time say come
  Headline: get sure want use
  Rouge Score: 0.1999999952000001
  Summary: get sure want use <EOS>

- **Article**: deep conditioning mask need egg white avocado tablespoon olive oil tablespoon honey measurements altered depending long thick hair using blender food processor add ingredients blend high creamy consistency formed blender mash ingredients together manually take bit time scoop mask onto fingertips spread throughout hair focusing primarily ends strand used mask wrap hair shower cap prevent mask rubbing clothes leave mask minutes allow time soak hair leave longer would like make sure wait least minutes remove shower cap run hair warm water shower using warm water wash remaining mask need use shampoo try avoid dry hair finished washing hair dry hair towel apply mask hair times week optimal results otherwise apply avocado mask hair whenever starts appear dull frizzy dry damaged touch

  Headline: washing hair dry hair towel apply mask hair times week optimal results
  Rouge Score: 0.27266186824905547
  Summary: use hair water hair hair hair times week optimal

Some Loss Plot:

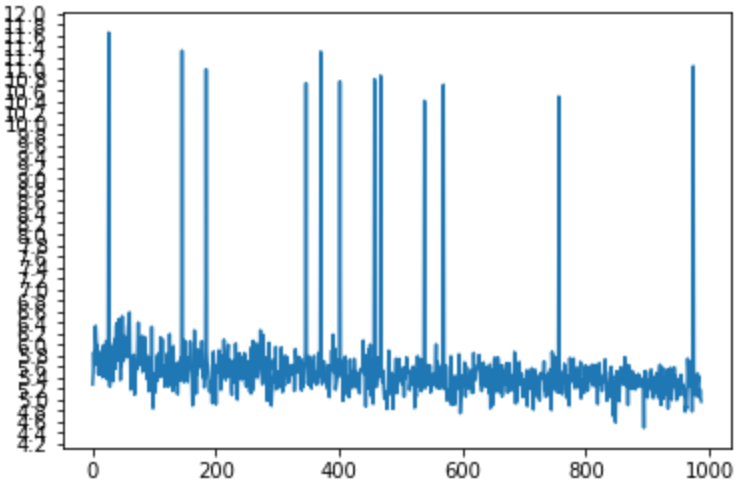

Fig 8:Loss Plot For Sl-6

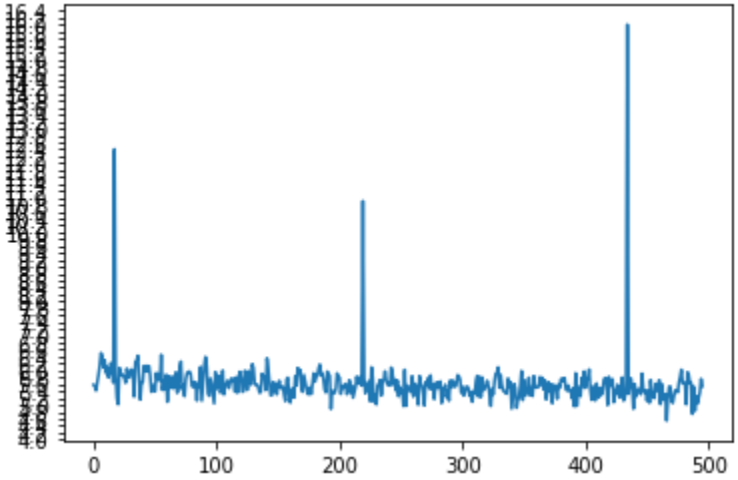

Fig 9: Loss Plot For Sl-2

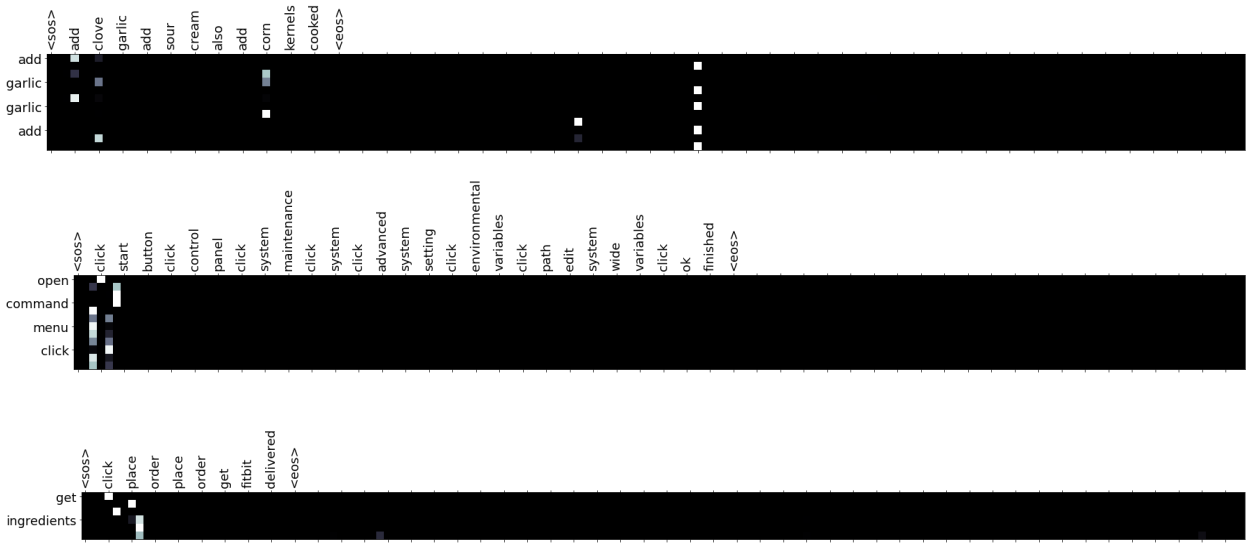

Fig 10: Work of attention in our System

6. Conclusion

Text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails and blogs. This is due to the fact that there is information overload in these areas, especially on the World Wide Web.Summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document .Both extractive and abstractive methods have been researched. Most summarization techniques are based on extractive methods.Abstractive summarization as of now requires heavy machinery for language generation and is difficult to replicate into the domain specific areas.

References

1. Einat Amitay and Cécile Paris. 2000. Automatically summarising web sites: is there a way around it?. In Proceedings of the ninth international conference on Information and knowledge management. ACM, 173–179
2. J-Y Delort, Bernadette Bouchon-Meunier, and Maria Rifqi. 2003. Enhanced web document summarization using hyperlinks. In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia. ACM, 208–215.
3. Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2007. Comments-oriented blog summarization by sentence extraction. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 901–904

4. Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 291–298
5. Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing microblogs automatically. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 685–688
6. https://arxiv.org/abs/1810.09305
7. https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/
8. https://towardsdatascience.com/text-summarization-from-scratch-using-encoder-decoder-network-with-attention-in-keras-5fa80d12710e