Extractive Multi document News summarizer using BERT

Project report submitted in fulfilment of the requirements

for the award of

Degree of Bachelor of Technology in

Computer Science and Engineering

by

Tarishi Jain (2016UCP1443)

Sumit Kumar (2016UCP1459)

Preet Yadav (2016UCP1436)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY

Certificate

We,

TARISHI JAIN (2016UCP1443)

SUMIT KUMAR (2016UCP1459)

PREET YADAV (2016UCP1436)

Declare that this thesis titled, “Extractive Multi document news summariser using BERT” and the work done in this project are of our own. We certify that

- This thesis work has been completed while pursuing the Degree of Bachelor of Technology in computer science and engineering stream at Malaviya National Institute of Technology ,Jaipur (MNIT).
- The part of the project which was earlier submitted for any qualification at MNIT is clearly stated in this thesis.
- Any help taken from word published by others and work done by others is clearly stated with the source.

Signed:

Date:

Dr. Yogesh Kumar Meena
Associate Professor

MAY 2020

Department of Computer Science and Engineering
Malaviya National Institute of Technology, Jaipur

Abstract

Multi-document text summarisation models can efficiently find informative sentences from a set of documents provided of same domain. But this type of summarisation comes with many challenges like complete extraction of information, ordering of sentences, removing and duplicate information, etc. There has been projects with various approaches with improving performance in this field.

In our project we attempt to do the same via new model implemented by Google in 2018 as a solution for various NLP problems: **BERT(Bidirectional Encoder Representations from Transformers)**, for extractive multi document summarization .BERT model is combined with unsupervised summarisation approach of k-means model of clustering , centroid-based model through compositionality of BERT embeddings, MMR and then finally ordering sentences chosen for the summary. Our method gave fairly good results when evaluated on ROUGE using DUC2002 dataset.

Acknowledgements

We would like to thank Dr Yogesh Kumar Meena Sir, Associate Professor, Department of CSE, MNIT, Jaipur for providing us this chance to work under his guidance on the final year B.Tech project. He constantly helped us from deciding fields and topics for the project to the ways of accomplishing the project with good efficiency and before the deadline.

Sumit Kumar (2016UCP1459)
Preet Yadav (2016UCP1436)
Tarishi Jain (2016UCP1443)

Contents

List of Figures

Chapter 1

Introduction

We can see various news articles publishing on the same news .For better understanding of the news, there is need to have a system which when given all such news articles pertaining to same headline, can generate effective and informative summary out of it. Multi-document news summarization made it possible to do so by taking out the gist of articles and providing it to the user for better understanding .

There are two popular ways of summarisation text. One is extractive summarisation, where summary is formed by taking sentences of original document, other form is abstractive in which the documents are first thoroughly read and then a whole new set of sentences are formed to be included in the summary. We chose extractive summarisation techniques.

Many extractive models have been proposed recently (Nallapati et al., 2017[4]; Dong et al., 2018[6]), but the performance of extractive summarization can still boost by the improvement on automatic metrics like ROUGE.

In this thesis, we focus on designing automatic summarization tool of multiple news articles using BERT to create sentences and word embeddings and then passing those embeddings to various extractive summarization models and shown the result on the DUC-2002 dataset. We had also build a user interface for end users to directly use our service. We had worked on tackling various add on problems that comes with Multi-document summarization like duplication of information and to ensure correct order of summary sentences. So we try to incorporate various methods along with BERT embeddings to find better results for summarizing multiple news articles.

Chapter 2

Important Terms and Concepts

2.1 Text summarization

Text summarization is the method for generating the relevant and important sentences from a relatively larger text, making it easier to analyse and understand without losing the overall meaning. There are two ways to achieve it, extractive and abstractive way.

2.1.1 Extractive summarization

Extractive text summarization aims at extracting sentences from original set of documents according to some determined features, and these features decide which sentences are more important and precedence over other sentences. By this process we determine important sentences and include them in summary.

2.1.2 Abstractive summarization

While in the case of extractive summarization, sentences remain same as they were in original text, In the case of sentences are changed but they still have the same meaning and deliver same information. This works at generating more human like summaries.

2.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained model developed and trained by Google. As BERT is already pretrained or a large dataset, it saves us a lots of efforts of training the neural network. BERT has further increased performance in NLP tasks which makes it more desirable to use in comparison to other models.

2.2.1 BERT embedding

It learns words from a large text corpus and design a vector form of representation of words given and captures its semantic relationship with other words present in its text corpus. BERT model takes the sentences in input and returns word embeddings in the output. BERT has better understanding of text as compared to other model because it reads the whole text at once instead of reading it from forward to backward or in one direction. Word embeddings are a vector representation of each sentence extracting features and representing them in terms of integer values.

2.2.2 Bert in summarization

BERT gives us output vectors , these vectors are represented in the form of tokens. Now BERT alone does not gives us summarization of text, It only gives us word embeddings of sentences. We further need to fine-tune BERT which means apply various algorithms on BERT to generate summary from our articles. We use various models for fine tuning BERT which are given below:

2.3 K-means Algorithm

In K-means algorithm, we divide all the data into a predefined number of clusters, each of these cluster contain some part of data. In the same cluster it tries to put the datapoints which share more similarity. One important factor in this algorithm is deciding the value of number of clusters. As having too large value or too small value of cluster results in poor results so we will have to decide optimum value of number of clusters using some technique.

2.4 Centroid-based through Compositionality of BERT Word Embedding [9]

In ordinary centroid-based method, tf.idf score is calculated of the vocabulary words and those which exceed a given threshold value, are taken as centroid words. The sentences which are comprising of most of the centroid words will be taken as part of summary.

This ordinary method, when combined with BERT word embeddings can prove to be very effective. In the modified centroid based method, upon calculating the centroid words as was in original method, the BERT word embeddings of the centroid words are stored in a table E, and centroid vector is calculated as the mean of embeddings of centroid words as given below:

$$C = \sum_{w \in D, tfidf(w) > t} E[idx(w)]$$

,where idx(w) is the function that returns the index of the word w in the vocabulary.

The sentence vector is calculated as the mean of centroid word embeddings which it contain and given by:

$$S_j = \sum_{w \in S_j} E[idx(w)]$$

The cosine similarity measure is then applied between centroid vector and sentence vector to predict the score of that sentence.

$$sim(C, S_j) = \frac{C^T \bullet S_j}{||C|| \cdot ||S_j||}$$

The sentences are then sorted in decreasing order of their score and top sentences are extracted for summary.

2.5 MMR (Maximal Marginal Relevance)

MMR is a model that is used to ensure that there is lesser redundancy in our final results. As we work on multi document summarization of news articles, many articles will have same information and that information may be selected into results so we need to remove redundancy before giving final results.

MMR model gives us two formulas to apply this model. The first formula is given below:

$$MMR \overset{def}{=} Arg \max_{D_i \in C \setminus S} \left[\lambda \left(Sim_1\left(D_i, Q\right) - \left(1 - \lambda\right) \max_{D_j \in S} Sim_2\left(D_i, D_j\right) \right) \right]$$

Where C represents all of the input documents, Di is the current sentence we are deciding whether to include in summary or not, Dj is the current state of summary, Q is the query to which we want or sentences in result have more similarity. We make query by choosing the best sentences in the text.

Another formula which is given below is used to find similarity between current summary and sentence on which we are deciding whether to include in summary:

$$Sim_1\left(u, v\right) = Sim_2\left(u, v\right) = \frac{\sum_{\mathrm{w} \in v} tf_{\mathrm{w},u} tf_{\mathrm{w},v} \left(idf_{\mathrm{w}}\right)^2}{\sqrt{\sum_{\mathrm{w} \in u} \left(tf_{\mathrm{w},u} idf_{\mathrm{w}}\right)^2}}$$

2.6 Sentence Ordering

We have used a simple greedy approach to order sentences in multi-document summarisation. The assumption taken here is that a good sentence ordering in summary implies that there should be good level of similarity between all adjacent sentences as repetition of word is one of the sign of text coherence. Coherence of summary S which comprises sentences from S1 to Sn is given by the following equation. For calculating Similarity(S_k,S_{k+1}), we have used the cosine similarity function.

$$Coherence(S)=\sum_{k=1}^{n-1} Similarity(Sk, Sk+1)$$

The sentence is placed in a document based on the greedy approach by calculating the coherence score. We select first sentence from the extracted summary and placed it in the ordered set S. Remaining extracted sentences are added in incremental way to the document set S to get the final order of summary sentences.

Chapter 3

Literature Survey and Related Work

To word and design a good model in summarization of news articles, It was necessary to get acquainted to previous advancements in this field, what kind of models and approach has already been used, advantages and limitations of each approach. Previous non-neural models in the summarization of multi documents have been developed in both fields: extractive (Erkan and Radev, 2004[11]) as well as abstractive (Ganesan et al., 2010[12]). Recently, a lot of focus has been given on neural methods in text summarization, but it is mainly used in single document summarization and in both approaches extractive approach [4] and abstractive methods (Cohan et al., 2018[13]) . Except these the paper (Alexander R. Fabbri and Tianwei She, 2019[14]) has further further improved the quality of summary and done a good job in the case of abstractive summarization of multi-news documents.

Early progress in the field of summarization of more than one documents was mainly with help of methods like K means clustering method and graph based techniques.

Centroid based method is another one of models which has emerged lately and has gained popularity due to it’s good results[9]. To remove redundancy, a proposed famous model is MMR model which was published lastely by Corbonell[10].

In addition we planned on using BERT [7], which is a pre-trained model developed by Google in 2018, that has given further good results in the field of NLP. Since BERT is a new model and it was released lately, there is still a lots of work that can be done using this model in the field of text summarization. While there are some papers who have done good work in text summarization using BERT model like (Derek Miller, Georgia Institute of Technology [2]) which uses BERT for extractive summarization and generating short summaries of lectures. Another paper written by (Yang Liu[16]) which uses many techniques to fine tune BERT model like classifier, Transformer and LSTM.

Along with BERT, we have applied Centroid-based through Compositionality of BERT Word Embeddings[9] paper’s approach of modifying the original centroid based approach and used BERT for word embedding in it.[17]

There is no work done using the BERT model in case of multi-document summarization. Our proposed project is the summarization of multi news documents which is introduced next.

Chapter4

Proposed Method

Our proposed model is proposed in two steps, First step is where we process our text and generate word embeddings with help of BERT. In second step many models are used on word embeddings generated by BERT model to give us good summarization results.

4.1 Text pre-processing:

- This module takes the text from input documents and cleans it. It generate sentences from large text corpus.
- These sentences are then passed to BERT model, which will give output sentence embeddings.
- BERT will take input sentences and convert them into a vector of floating numbers of fixed size, even if the sentences are of varying length, thus producing fixed size sentence embeddings.

4.2 Summarization models

In this step we apply various models on the sentence embeddings given by BERT to extract the most useful sentences and generate summary:

4.2.1 k-means:

- The unsupervised k-means algorithm takes as input ,the embedding sentence vectors created by BERT and combine these sentences into clusters. From each cluster, the sentence which is closest to the centre of that cluster is taken into summary..
- In this model, deciding the number of clusters is a challenge. Choosing a

very big value of No of clusters or very small value of clusters, both will result in bad summary quality.

4.2.2: Centroid based BERT embedding model:

- The tf-idf scores are calculated for the vocabulary of the input given and top scores words are taken and called as most important words.
- The centroid vector is then calculated as the mean of BERT word embeddings of these most important words.
- The sentence embeddings are calculated as mean of embeddings of the important words they contain.
- The sentences having high cosine similarity score with the centroid vector are taken into results.

4.2.3 MMR and sentence position:

- The MMR model is used to ensure that there are no duplicate sentences or sentences having redundant information.
- In last stage, coherence of the summary is calculated by calculating the similarity among sentences and the sentences are sequenced in a way so that adjacent sentences have higher relative similarity thus producing good coherence of summary, to put them in correct order in the summary.

In the end the summary will be evaluated by using ROUGE measures of evaluation.

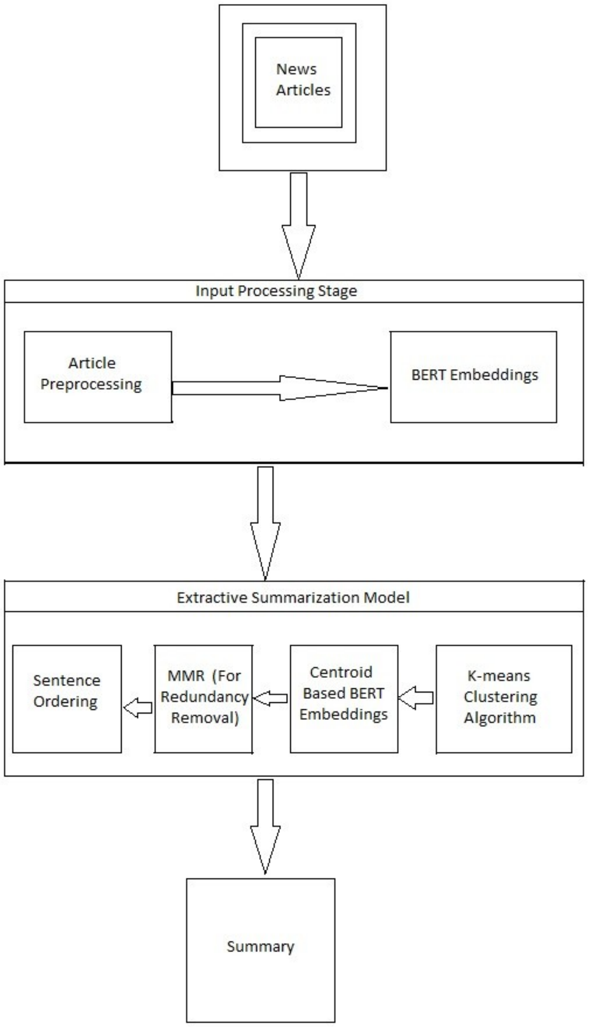
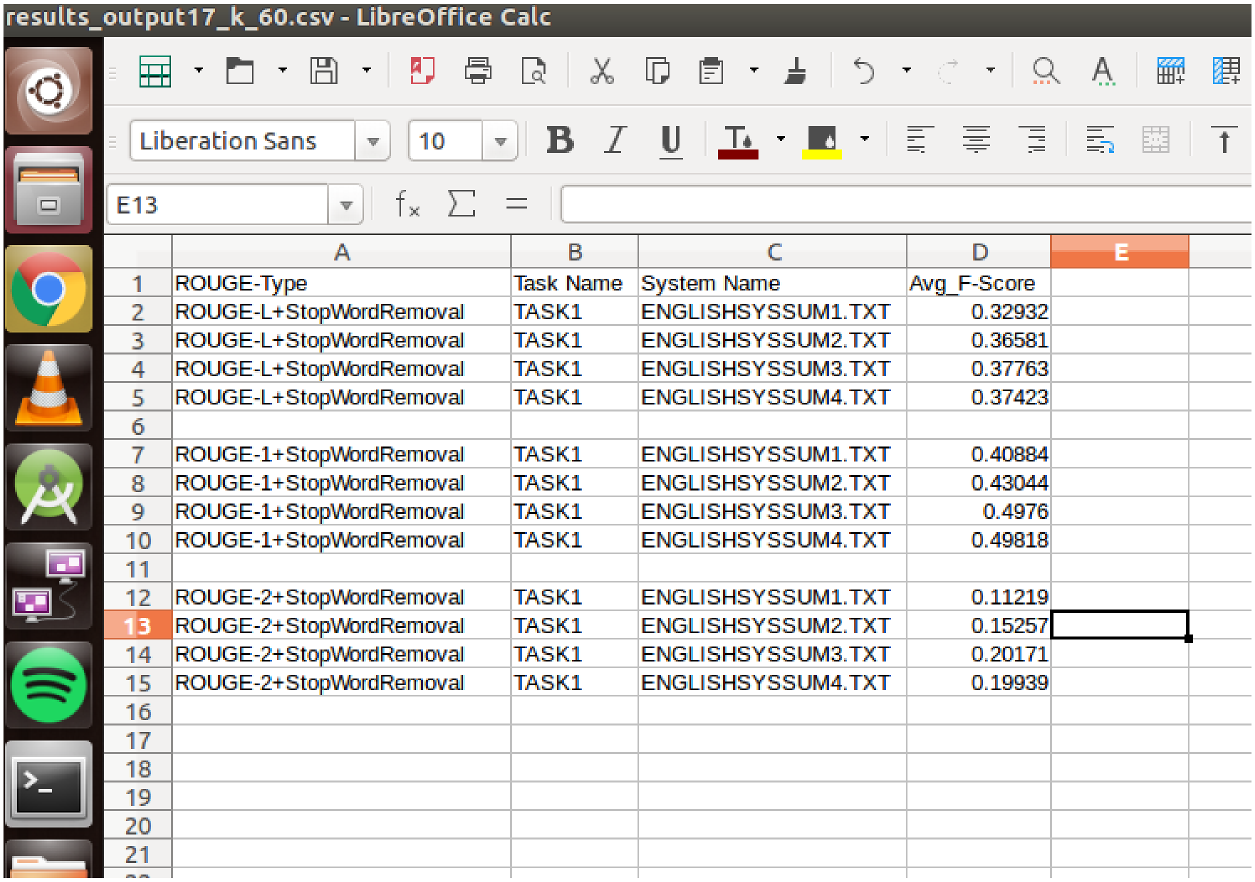

News Articles
Input Processing Stage
Article Preprocessing
BERT Embeddings
Extractive Summarization Model
Sentence Ordering
MMR (For Redundancy Removal)
Centroid Based BERT Embeddings
K-means Clustering Algorithm
Summary

Figure:Architecture for summarization model

Chapter 5

Experimental Results

We measured our results with the help of ROUGE matrix on DUC2002 dataset. The ROUGE value of resultant summary of every stage is compared with the reference summary provided in the dataset. It can be easily seen from below figure that the ROUGE value keep on improving as we add other stages to the model.



	A	B	C	D	E
1	ROUGE-Type	Task Name	System Name	Avg_F-Score	
2	ROUGE-L+StopWordRemoval	TASK1	ENGLISHSYSSUM1.TXT	0.32932	
3	ROUGE-L+StopWordRemoval	TASK1	ENGLISHSYSSUM2.TXT	0.36581	
4	ROUGE-L+StopWordRemoval	TASK1	ENGLISHSYSSUM3.TXT	0.37763	
5	ROUGE-L+StopWordRemoval	TASK1	ENGLISHSYSSUM4.TXT	0.37423	
6					
7	ROUGE-1+StopWordRemoval	TASK1	ENGLISHSYSSUM1.TXT	0.40884	
8	ROUGE-1+StopWordRemoval	TASK1	ENGLISHSYSSUM2.TXT	0.43044	
9	ROUGE-1+StopWordRemoval	TASK1	ENGLISHSYSSUM3.TXT	0.4976	
10	ROUGE-1+StopWordRemoval	TASK1	ENGLISHSYSSUM4.TXT	0.49818	
11					
12	ROUGE-2+StopWordRemoval	TASK1	ENGLISHSYSSUM1.TXT	0.11219	
13	ROUGE-2+StopWordRemoval	TASK1	ENGLISHSYSSUM2.TXT	0.15257	
14	ROUGE-2+StopWordRemoval	TASK1	ENGLISHSYSSUM3.TXT	0.20171	
15	ROUGE-2+StopWordRemoval	TASK1	ENGLISHSYSSUM4.TXT	0.19939	
16					
17					
18					
19					
20					
21					

Figure: Summary quality measured on ROUGE matrix on DUC2002 dataset

Summary named ‘ENGLISHSYSSUM1’ is the summary generated by applying K-means clustering method on top of BERT model. We generate summary ‘ENGLISHSYSSUM2’ by making use of K-means+Centroid based BERT embedding algorithm which removes some of sentences which were included in K-means model but they contain poor information.

After this, we also add MMR model to our project and summay after this stage is named as ‘ENGLISHSYSSUM3’. MMR model is used to remove redundancy from the summary. As we can see that after applying MMR algorithm, redundancy is removed and the ROUGE score of the matrix improves showing that the quality of our summary also improves.

After this, in the last step we add Sentence Ordering algorithm to our project, as this algorithm only changes the sequence of sentences, It doesn’t add or remove any sentences so quality doesn’t change much here.

A comparison is also shown in Table1 between other multi document extractive summarisation models and our proposed model outperforms them to a good extent.

System	**ROUGE**-1	**ROUGE**-2	**ROUGE**-L
LSA [8]	37. 92	7.74	35.02
LDA [8]	35.69	6.26	32.71
Random [8]	32.028	5.432	29.127
Lead [8]	31.446	6.151	26.575
DSDR-non [8]	39.562	7.440	35.345
Single layer[18]	37.75	8.82	-
C SKIP [9]	38.81	9.97	-
PG-MMR [19]	40.44	14.93	-
Our model	48.9	18.7	37.25

Table 1: Comparison of results on rouge matrix with other popular models.

Model	**ROUGE-1 Score**
DUC-best [20]	0.4986
BSTM [20]	0.48812
FGB [20]	0.48507
LexPR [20]	0.47963
Multilayer [20]	0.4215
Based on graph independent sets[21]	0.58060

Our model	0.4893

Table 2 : shows some of the model’s ROUGE results when run on DUC-2002 dataset.

In this model, one factor that affects output summary results a lot is the number of clusters in K-means algorithm. Choosing a very big value of No of clusters or very small value of clusters, both will result in bad summary quality. We represent No. of clusters by k. To choose the value of k, we run the model on various datasets of DUC2002 data with different value of k. We plot a graph between no of clusters and summary quality measured on rouge matrix. The graphs plotted are shown in figure below:

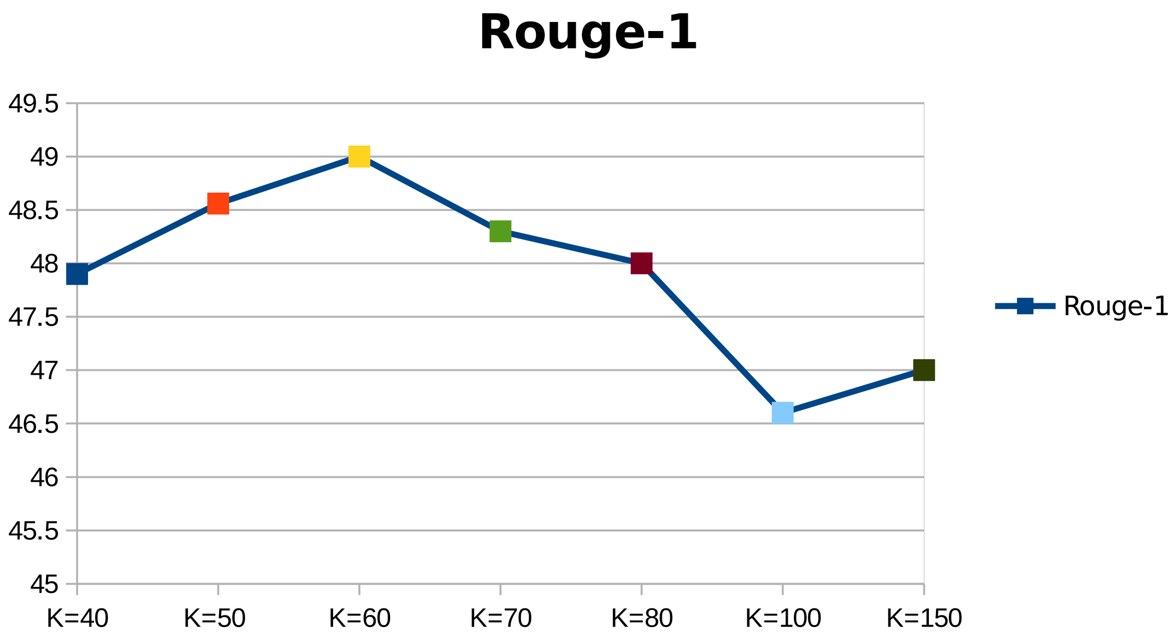


Figure: Rouge-1 results on varying value of clusters in K means model

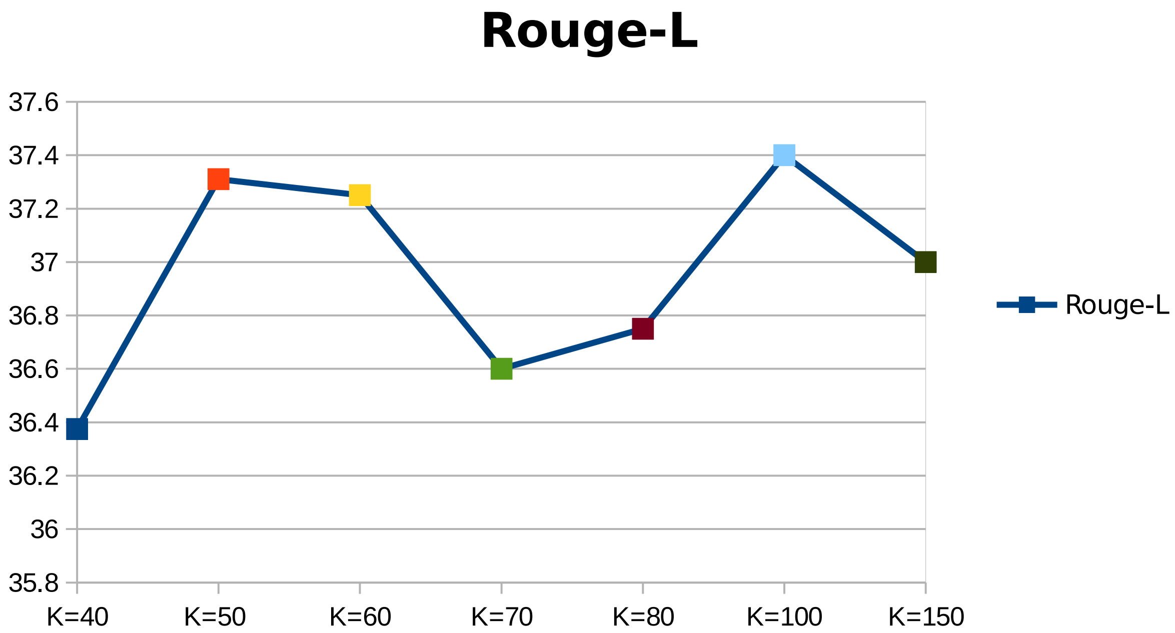


Figure: Rouge-L results on varying value of clusters in K means model

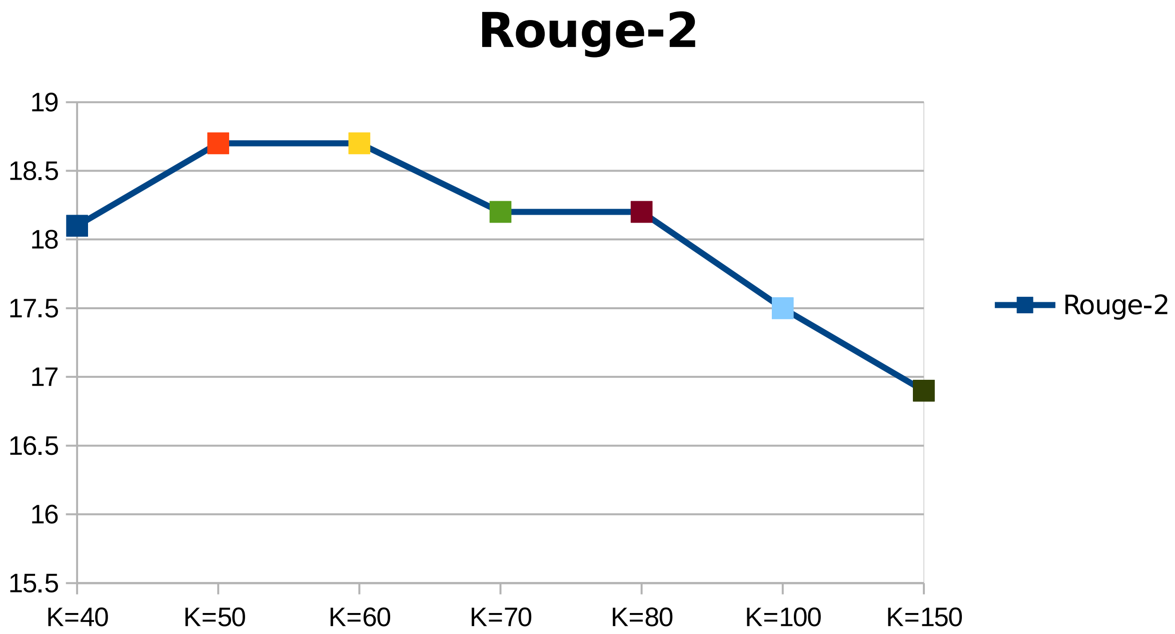


Figure: Rouge-2 results on varying value of clusters in K means

Now as shown in the graph plotted above, best value for number of clusters comes out to be 60 on DUC2002 dataset. The output summary length is set as 20 sentences in the case of DUC2002 dataset. But summary length can vary and then we will have to change the number of clusters again. So it is necessary to find a relation between the number of clusters and summary length. For this we also run our model on DUC 2007 dataset. Graph plotted is shown below:

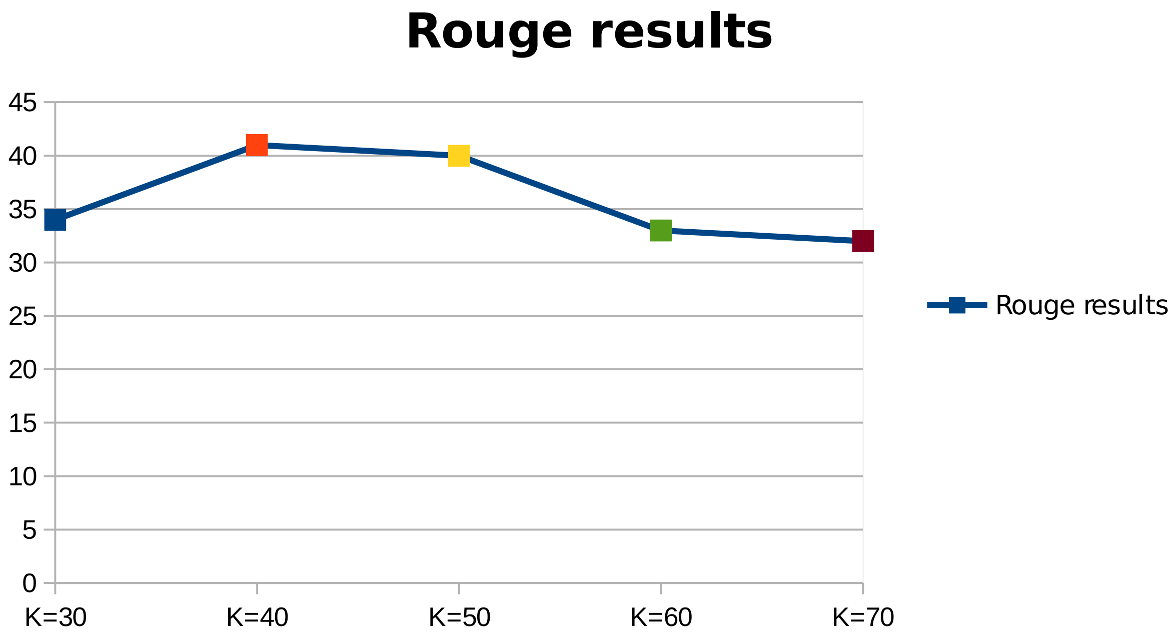
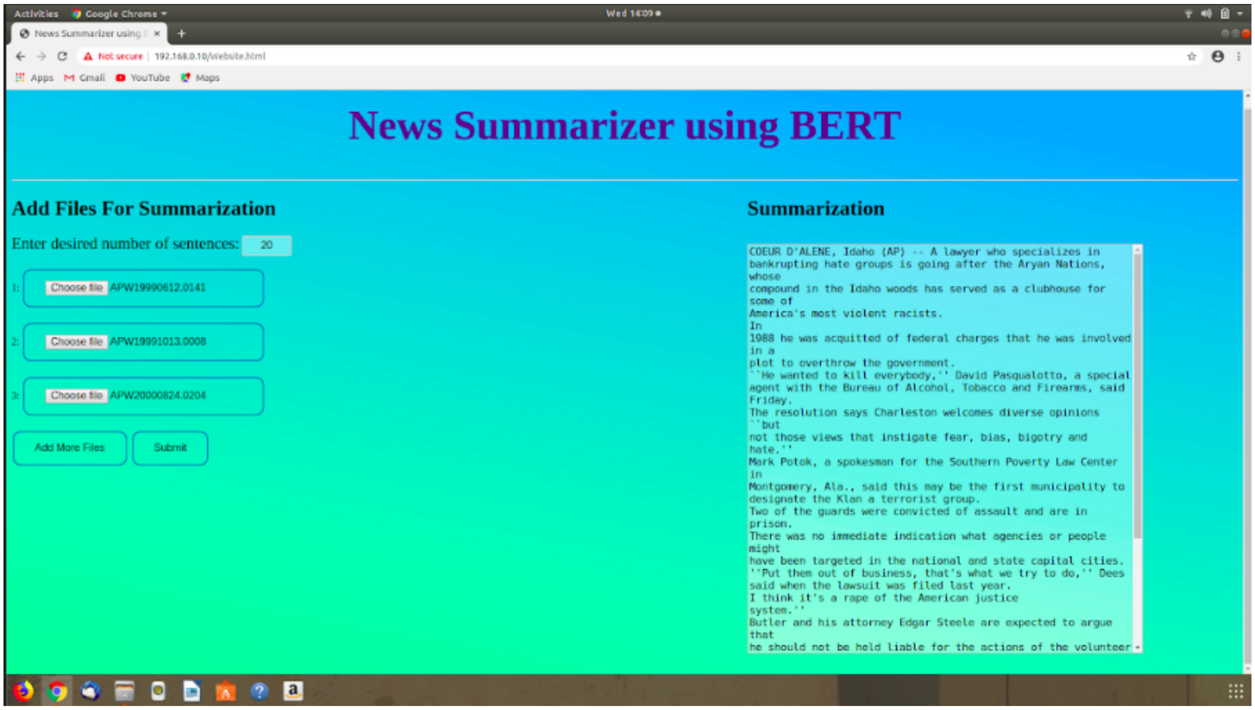


Figure: Rouge matrix results on DUC2007 dataset with varying Number of clusters

In case of DUC2007 dataset output summary length was set as 13 sentences and we get relatively good results at k=40.

So we set the number of clusters equal to the factor * summary length. Where factor determined from above 2 experimental results is taken factor=3.
Thus,

k=3*summary length(in sentences)



News Summarizer using BERT

Add Files For Summarization

Enter desired number of sentences: 20

1: Choose file APW19990612.0141

2: Choose file APW19991013.0008

3: Choose file APW20000824.0204

Add More Files   Submit

Summarization

```
COEUR D'ALENE, Idaho (AP) -- A lawyer who specializes in
bankrupting hate groups is going after the Aryan Nations,
whose
compound in the Idaho woods has served as a clubhouse for
some of
America's most violent racists.
In
1988 he was acquitted of federal charges that he was involved
in a
plot to overthrow the government.
``He wanted to kill everybody,'' David Pasqualotto, a special
agent with the Bureau of Alcohol, Tobacco and Firearms, said
Friday.
The resolution says Charleston welcomes diverse opinions
``but
not those views that instigate fear, bias, bigotry and
hate.''
Mark Potok, a spokesman for the Southern Poverty Law Center
in
Montgomery, Ala., said this may be the first municipality to
designate the Klan a terrorist group.
Two of the guards were convicted of assault and are in
prison.
There was no immediate indication what agencies or people
might
have been targeted in the national and state capital cities.
''Put them out of business, that's what we try to do,'' Dees
said when the lawsuit was filed last year.
I think it's a rape of the American justice
system.''
Butler and his attorney Edgar Steele are expected to argue
that
he should not be held liable for the actions of the volunteer
```

Figure: User Interface design for user end

In the end for easy access for user, we designed a web page as a User Interface for easy access to user.

Chapter 6

Conclusion

In this work, we have implemented BERT text embeddings alongwith K-means clustering ,centroid-based model using BERT embeddings, MMR, and sentence ordering to produce an extractive multi article news summariser system. Rouge results implemented on DUC2002 dataset implies that our proposed model is giving good results than various other multi document summarisation systems. Thus we can say that using BERT for text embeddings improves the quality of the summary for multi-document summarization problem to a good extent.

References

[2]: Derek Miller,Georgia Institute of Technology.Leveraging BERT for Extractive Text Summarization on Lectures

[4]:Ramesh Nallapati, Feifei Zhai, 2017. Summarunner: A recurrent neural network for extractive summarization of docu- ments. AAAI Conference.

[6]:Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarisation as a contextual bandit. In Proceedings of the EMNLP Conference.

[7]:Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

[8]: Hai Cao Manh, Huong Le Thanh, Tuan Luu Minh. 2019. Extractive Multi document Summarization using K-means, Centroid-based Method, MMR, and Sentence Position. In SoICT ’19:

[9]: Gaetano Rossiello, Pierpaolo Basile, Giovanni Semeraro. Centroid-based Text Summarization through Compositionality of Word Embeddings, April 2017.

[10]:Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Research and Development in Information Retrieval.

[11]:Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization. Journal of artificial intelligence re- search, 22:457–479.

[12]:Kavita Ganesan, ChengXiang Zhhai, 2010. Opinosis: A Graph Based Approach to Abstractive Summarization.2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, , Beijing, China, pages 340–348.

[13]:Arman Cohan, Franck Dernoncourt,Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pages 615–621.

[14]:Alexander R. Fabbri,Tianwei She,Department of Computer Science,Yale University Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model(2019)

[15]:Ashish Vaswani, Noam Shazeeer, Niki Parmar, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

[16]:Yang Liu,Institute for Language, Cognition School of Informatics, University of Edinburgh,10 Crichton Street, Edinburgh E8 9AB, Fine-tune BERT for Extractive Summarization.

[17]: Mir Tafseer Nayeem, Yllias Chali, University of Lethbridge Lethbridge, AB, Canada, Extract with Order for Coherent Multi-Document Summarization.

[18]: Lu´ıs Marujo , Ricardo Ribeiro.Extending a Single-Document Summarizer to Multi-Document: a Hierarchical Approach,2015

[19]: Logan Lebanoff, Kaiqiang Song and Fei Liu ,2018.Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization

[20]: Jorge V. Tohalino, Diego R. Amancio,2017.Extractive Multi-document Summarization Using Multilayer Networks
[21] Taner Uçkan, Ali Karcı ,2020.Extractive multi-document text summarization based on graph independent sets. Egyptian Informatics Journal