2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Analytical study of Text Summarization Techniques

Dr. Pooja Raundale
Prof. and Head, MCA,
Sardar Patel Institute of Technology
pooja@spit.ac.in

Himanshu Shekhar
Student, MCA,
Sardar Patel Institute of Technology
himanshu.shekhar@spit.ac.in

***Abstract*—Summarization of Text is extracting important information from a body of text and present it in the form of a concise summary. The need for summarization has increased in recent times. The importance of having a simple concise summary of information in the news, business and research domains is paramount. Automatic text summarization is a well known task in the NLP (Natural Language Processing) field. Text Summarization techniques can be large divided into two groups: Extractive summarization and Abstractive summarization. Extractive summarization is based on identifying key sentences or phrases from the source text and grouping them to produce a summary without rewriting or paraphrasing the original text. Abstractive summarization is based on utilizing a deeper understanding of the source text and generating new sentences, not present in the original text, which improves the summary by reducing redundancy and focusing on the meaning of the source text.**

In this study we implement and compare the performance of various automatic summarization methods in order to gain insight into how long the methods take to implement and how accurate and human-like the generated summaries are. We aim to learn the pros and cons of the various techniques used by utilizing summary scoring as well as manual inspection of generated summary.

Keywords— Automatic Text Summarization, Natural Language Processing, Abstractive Summarization, Extractive Summarization

I. Introduction

Recently, there has been a surge in the amount of text data from various sources. This volume of text is a valuable source of expertise and facts that must be efficiently summarised in order to be useful. The main goal of this issue is to automate text summarization. People are overwhelmed by the vast amount of online knowledge and documents as the Internet has grown dramatically. Because of the growing availability of documents, extensive research into automated text summarization is needed. There is a lot of study being done these days on text summarization. These kinds of things are becoming more common as the amount of knowledge on the internet grows.

II. Literature Review

The method of extracting or compiling relevant information from an original text and providing it in the form of a description is known as text summarization. Many applications, such as search engines, business evaluations, and industry assessments, now need text summarization. Summarization allows you to get the details you need in less time. This paper attempts to outline and present the history of text summarization from its inception to the present. The two main approaches to summarization, extractive and abstractive, are explored in depth. The summarization techniques used vary from structured to linguistic. This information presents an abstract view of the current text summarization research scenario. We have decided to apply a few of the techniques after studying and investigating the different techniques, both extractive and abstractive. The methods that were selected were as follows.

TF-IDF is a numerical measure used to measure the importance of a word in a document according to how often it was contained in that document and in any collection of documents. Instead, it is a short term frequency-inverse document. The idea behind this measure is that, if a word is often found in a document, it should be important and that word should be highly valuable. However, if a word appears in too many other documents, it may not be a unique identifier, so we should assign it a lower score. This is used in text summarization, as we can write words based on TF-IDF metrics and then use the extractive summarization approach as the top scoring phrases for summary.

TextRank[5] is a graph-based ranking model. The approach is based on the PageRank algorithm of Google. We first need to build a text-related graph to use TextRank for automated summary, where the graph vertices are representative for the units to graded. In order to extract sentences, the goal is to grade whole phrases and a vertex for each sentence in the text is therefore added to the graph. We define a different relationship which determines the connection between two phrases if there is a relationship ”similarity.” Similarity is measured as a function of the overlap of their contents.

Seq2Seq[10] is the machine translation method based on encoder decoder that maps a sequence input to a sequence output with a tag and attention value. The model uses a ”Attention”[1] technique, which enables the model to focus on various parts of the input sequence in each stage of the output sequence to preserve the context from start to finish. The encoder encodes the phrase word for word with an index of vocabulary or known words with index and predicts the output of the encoded entry by sequentially decryption the input and, if possible, attempts to use the last entry as the next one. In this way, the next input for creating a sentence can also be predicted. A token is assigned to mark the end of the sequence for every sentence. There is also a token to

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544804

mark the end of the output at the end of the prediction. Thus, it passes from the encoder to the decoder to generate the output.

Seq2Seq has two main disadvantages. The model is often able to repeat phrases and it can't understand words that are not in it's vocabulary. The Pointer generator[9] approach resolves the problem of out of vocabulary words and other imprecision in previous technique's summary output. A hybrid abstractive and extractive uses technique generating new (abstractive) words from the semantic context, or copying them from source text (extractive). This allows the model to create new words and phrases. If a word is out of vocabulary or where it cannot find a semantic meaning, then the model points to the word in source and copies it into the generated output summary. There is a generational chance for each word to indicate to the model whether the source word should be paraphrased or the source word itself should be copied. It also uses a "coverage" technique where there is introduced a coverage loss that punishes that model if words or phrases are repeated in the output summary. The problem of repetition is addressed here.

III. Methodology

The proposed goal of this research article is to implement various algorithmic and deep learning techniques and analyze and compare the performance and accuracy of the techniques. Breakdown of the steps taken to achieve the goal-

A. Understanding the Problem domain

Text Summarization is a problem with a broad area and in the last decade a lot of research has been done. We aimed to read and understand key research papers to understand the different techniques and approaches employed to address the problem.

B. Choosing data set

Second step was to analyze and choose a training data set to utilize with every technique.
There are two widely used data sets in the domain of automatic text summarization.

1) **CNN / Daily Mail[3]:** The dataset contains online news articles (average 56 tokens). The version processed includes 287,226 training, 13,368 test and 11,490 test articles.

2) **GIGAWORD[8]:** It contains very short input documents (average 31.4 tokens) and summaries (average 8.3 tokens). It includes 3.8M training, 189K and 1951 test articles.

We have chosen to use CNN/Daily Mail dataset in order to achieve consistent ROUGE scoring.

C. Dataset Processing

The dataset is a text file with source text and a handwritten summary of the information collected from both the CNN and the Daily Mail websites. We tokenized and saved the sentences in text files. Tokenized data is then transformed into bin and vocab files. The stories are tokenized, read from the files, lower cased and written to binary files. These binary files are then divided to improve memory and reading performance (1000 examples per chunk). The vocab file generated works as the vocabulary for the model. Our data set contains 52,000 words for the vocabulary. This is the last step in the processing of data before we start training the various models.

D. Model Implementation

All four techniques chosen were implemented on the Google Colaboratory platform. Due to performance considerations and the ease of sharing or publication of the implementations, we chose to implement the code on this platform. The Notebook instance was powered by an Nvidia Xeon 2.30GHz CPU and an Nvidia K80 GPU 12GB. The techniques were applied using Tensorflow and Pytorch in Python 3. Dataset was stored on Google Drive and then read from within the Google Colaboratory platform by mounting the respective folders. Both the extractive techniques, TF-IDF and TextRank, use algorithmic approach and therefore were implemented without any training. The abstractive techniques, Seq2Seq with Attn and Pointer-Generator, were trained till the running average loss via exponential decay was approximately around 1. In both the abstractive models we used Glove pre-trained vectors to initialize word embedding.

E. Evaluating the techniques based on Rouge metrics

Both models have been tested using the summaries generated from the test set and ROUGE metrics have then been calculated. The name ROUGE[4] is for Gisting Gisting Recall-Oriented Understudy, a set of method and software packages that is used for evaluation in the natural language processing field of automated summary software. The metrics automatically compare the summary produced with a (human) summary of a reference(s). Rouge metrics compare the reference summary and the summary to be evaluated.

1) ROUGE-1 refers to the unigram overlap of each word between the reference and generated summaries.
2) ROUGE-2 refers to the bigrams overlap of bigrams between reference and generated summaries.
3) ROUGE-L: It is based on Longest Common Subsequence. ROUGE-L takes into account sentence level structure similarity and identifies longest co-occurring sequence n-grams.

For all the techniques, the mean scores were taken. We also scored the first three lines of the dataset news articles on the given ROUGE metrics (Lead 3).

IV. Analysis

The generated summaries and the ROUGE evaluation metrics can be used to analyze and compare the various techniques.

A. Rouge Scores

Technique	ROUGE-1	ROUGE-2	ROUGE-L
TF-IDF	35.2	17.39	25.23
TextRank	39.83	17.62	19.48
Seq2Seq with Attention	31.33	15.66	33.42
Pointer Generator with Coverage	39.53	17.28	36.38
Lead 3 baseline	40.34	17.70	36.57

1: Rouge Scores

From the tabulated scores in the above table we can infer the following-

- Extractive techniques (TF-IDF and TextRank) achieve very high scores for ROUGE evaluation. This maybe because ROUGE and other automatic metrics rely on lexical overlap which favors extractive techniques because they simply extract sentences from the source text to include in the summary.
- Extractive techniques achieve very high scores despite the fact that they are algorithmic approaches and thus much simpler in implementation.
- The earlier model in abstractive summarization (Seq2Seq with Attention) scores a lot lower than the other models. This score is massively improved by the later approach (Pointer Generator with Coverage) which builds upon the earlier technique and solves the two main problems of errors and repetition.
- The first three lines of a source text (Lead 3) ends up being a really good summary for the whole text when it comes to news articles. This maybe because online news articles are written in a way where the opening sentences are meant to be the words that provides the gist of the article and draws the reader to read more of the text.

B. Generated Summaries

There are some drawbacks to ROUGE evaluation and other such automatic evaluation metrics even though they are widely used in the research community in this type of problem.

- They only evaluate the selection content and do not take other quality factors into account, such as fluency, grammar, consistency, etc.
- They rely mainly on lexical overlap for the evaluation of content selection, although the same content as the reference without any lexical overlap can be expressed by an abstract summary.
- Because of the subjectivity of the summary and the relatively low agreement between annotators, the measures have been designed to be used with several reference summaries per input. However, only a single reference can be found in recent datasets like CNN/Daily Mail and Gigaword.

Original Text (truncated): lagos, nigeria (cnn) a day after winning nige-ria's presidency, muhammadu buhari told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. buhari said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the ter-rorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. buhari defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.
TF-IDF: "I promised the country free and fair elections. I have kept my word." Buhari, 72, will be sworn in on May 29. Who is Nigerian President-elect Muhammadu Buhari? Who is Nigeria's Muhammadu Buhari? Buhari praised voters for exercising their right peacefully.His campaign was also fiercely anti-corruption. Even the presidential vote had to be postponed because of the radical militants. Yet the violence persisted. Jonathan had been criticized for not doing enough to combat Boko Haram. The economy is another major issue. It also hosts many international oil companies and workers.
TextRank: "A new day and a new Nigeria are upon us," Buhari said after his win Tuesday. A day after winning Nigeria's presidency, Muhammadu Buhari told CNN's Christiane Amanpour that he plans to aggressively fight corruption that has long plagued Nigeria and go after the root of the nation's unrest.Before the election, African affairs analyst Ayo Johnson said the vote would come down to who could make Nigeria feel safe.
Baseline Seq2Seq + Attention: UNK UNK says his administration is confi-dent it will be able to destabilize nigeria's economy. UNK says his admin-istration is confident it will be able to thwart criminals and other nigerians. he says the country has long nigeria and nigeria's economy.
Pointer-Gen + Coverage: muhammadu buhari says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long his-tory of military rule, coups and botched attempts at democracy in africa's most populous nation.

2: Generated Summaries

Therefore, along with analyzing state-of-the-art automatic metrics like ROUGE, we also carried out a manual comparison of generated summaries. The above table gives an example of a news article and the generated summaries from the four implemented techniques.

From comparing the generated summaries in the above table we can infer the following-

- Extractive techniques like TF-IDF and TextRank generate summaries that appear to be more robotic and algorithmic rather than human-like. This is because they do not generate new sentences and simple extract high priority sentences from the source text to be represented in the summary. This means that paragraph structure and consecutive sentence relation and coherence is not taken into account which makes the summary seem robotic.
- Abstractive techniques like Seq2Seq with Attention and Pointer-Generator generate human like summaries that appear to be handwritten. We can see that Abstractive summarization is more advanced and closer to human-like interpretation.
- Seq2Seq with Attention model is a purely abstractive model and thus it cannot make sense of words that

are not in it's vocabulary (Eg. Name of President is an uncommon proper noun that does not exist in the training dataset). This generates lots of unknown keys (UNK) in the generated summary that need to be rectified manually. It also tends to repeat key points (Eg. First and Second sentence in the generated summary)

- Pointer-Generator being built on top of the previous model, utilizes a hybrid approach and coverage mechanism to produce better results. It copies the unknown words from the source text and implements a coverage loss that forces the model to reduce repetition in the generated summary. Thus it produces a human like summary with no errors and minimum repetition.

V. Conclusion

Looking at the generated summaries and the evaluation of the ROUGE scores we can conclude the following points.

1) Extractive Techniques score very well for ROUGE relative to the time to implement and process.
2) Seq2Seq with attention has no way of dealing with words not in it's vocabulary.
3) Pointer Generator with coverage model scores very well and deals with the two issues of previous deep learning models by reducing repetition and errors.
4) Lead 3 baseline sentences score very well because first three lines of a news article are coincidentally a very accurate summary of the whole article.

Thus it can be concluded that with every new technique the ROUGE scores have been gradually increasing. Extractive summarization techniques inherently score better in ROUGE because they are simple extracting sentences from the source text but they appear to be robotic and machine generated. Abstractive techniques score well and produce natural human like summary but they have a huge cost in training and require substantial dataset. In case of news articles, the first three sentences proves to be a very accurate summary of the whole article as indicated by the highest ROUGE scores out of all the techniques.

VI. Future Enhancements

Include newer Transformer based approaches to the problem.

Acknowledgment

The Sardar Patel Institute of Technology supported this research paper. We thank the faculty members whose comments have greatly affected the quality of the work.

References

[1] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2016.
[2] Cho, K. Deepmind Q&A dataset.
[3] Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend, 2015.
[4] Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 74–81.
[5] Mihalcea, R., and Tarau, P. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 404–411.
[6] Nallapati, R., Zhou, B., dos santos, C. N., Gulcehre, C., and Xiang, B. Abstractive text summarization using sequence-to-sequence rnns and beyond, 2016.
[7] NLP-progress. Tracking progress in natural language processing.
[8] Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization, 2015.
[9] See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks, 2017.
[10] Shi, T., Keneshloo, Y., Ramakrishnan, N., and Reddy, C. K. Neural abstractive text summarization with sequence-to-sequence models, 2020.