JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text Summarization using LED for MoM

1st Thiruvaazhi U
HoD, *Department of ISE*,
Kumaraguru College of Technology,
Coimbatore, India.
thiruvaazhi.ise@kct.ac.in

2nd Pawankumar S
UG Scholar, *Department of ISE*,
Kumaraguru College of Technology,
Coimbatore, India.
pawankumar.18is@kct.ac.in

3rd Guruprasath M
UG Scholar, *Department of ISE*,
Kumaraguru College of Technology,
Coimbatore, India.
guruprasath.18is@kct.ac.in

4th Jayaprakash J
UG Scholar, *Department of ISE*,
Kumaraguru College of Technology,
Coimbatore, India.
jayaprakash.18is@kct.ac.in

***Abstract*—Minutes of Meeting are a part of practically every professional meeting that allows for the tracking and management of workflow, as well as the gathering of the gist of the meeting material for attendees and non-attendees who may need the information to act on. Until now, taking MoM has been done manually, with one person listening to every conversation and taking notes on what needs to be stored. By automating this procedure, any company or management may have its MoM companion staged whenever and wherever they need it, without relying on humans.**

Keywords*— *MoM automation

I. INTRODUCTION

Our project's main goal is to employ NLP techniques to help automate the MoM process (Minutes of Meeting). Increasing the model's viability across domains by obtaining more labeled data and favoring the supervised approach. Increasing model knowledge using a variety of text data in order to obtain more meaningful summaries on many elements. The solution mechanism has been divided into three steps, according to the reverse engineering process, which was done to be more clear on the format of input our system receives and the required pre-processing to fine tune our mechanism for the summarization model, the three phases of solutioning are : Speech recognition is the process of recognising a person's voice and converting it into text. Speaker recognition is used to deal with inaccuracies in summary caused by people. Finally, summary entails not compromising on critical information and summarizing. Speech recognition is aided by high-performance, dependable APIs, such as Google's automatic speech recognition technology. Speaker diarization is a concept that identifies the speaker who speaks a specific sentence so that summarization can include speaker data. Finally, networks powered by bi-directional Long Short Term Memory aid summarization by understanding contexts from previous and subsequent sentences.

II. PROPOSED SYSTEM

A. Text Summarization

According to our literature review, text summarization focuses on capturing the target chunk of text and performing abstractive summarization, which is found to be a better style of summarization than extractive summarization. Abstractive-summarization focuses on comprehending the context of the input text and wording out its own phrases, compressing content but not delivering it. The goal of extractive summarization is to grasp the text and extract the most essential sentences from the input.

B. Issues identified

1) To handle a wide range of circumstances in summarization by training with a variety of texts and contexts, ensuring that the model is adaptable and consistent.
2) Reduce the amount of trainable parameters to simplify the model and reduce processing resources while maintaining performance.

3) Increasing the model's vocabulary and selecting appropriate embedding formats (tensorflow embedder, word2vec, etc.) and neural cells (LSTM, Bi-directional LSTM, RNN, and so on).
4) Making the model capable of training on big pieces of text and dynamically producing variable-length summaries.

III. METHODOLOGY

Let us take a look at the methodologies tried out for summarizing in this task. And focus more on LED based summarization.

A. Extractive summarization Seq2Seq model :

When it comes to text summarization, sequence to sequence encoder-decoder models are the gold standard. The encoder segment is in charge of bringing in contextuality vectors with word embeddings, which are then passed to the decoder segment, where it produces a summary. It takes a text paragraph as input and produces a summarized version of the same text. Its main hyperparameters are latent dimension, input text maximum length, summarization maximum length, padding type,activation, loss, optimizer functions.

B. Abstractive summarization using HuggingFaceTransformer and PEGASUS:

PEGASUS is a huge Transformer-based encoder-decoder model that has been trained on big text datasets such as CNN News, BBC News, and others. It's used to experiment with abstractive summarization. The transformer model integrates self-awareness and applies diverse influences to parts of input data, which are not always in order. This makes abstract summarizing easier. Similarly to Extractive summarization, the key sentences from the input document are screened/removed, and a single output sequence is formed from the remaining phrases. The best PEGASUS model is evaluated against a set of 12 downstream summarizing tasks, including emails, patents, stories, research, and so on. Furthermore, the model outperformed previous state-of-the-art results on low-resource summarization (i.e., 1000 instances vs. tens of thousands of training data), outperforming previous state-of-the-art findings. These results were finally validated through human evaluation, and it was able to accomplish human-level abstractive summarization on a variety of datasets.

C. Longformer Encoder Decoder

Due to its self-attention operation, which scales quadratically with sequence length, transformer-based models are unable to process large sequences. To overcome this issue, the Longformer has an attention mechanism that scales linearly with sequence length, allowing it to process documents with tens of thousands of tokens or more. The attention mechanism in Longformer is a drop-in substitute for ordinary self-attention, combining local windowed attention with task-motivated global attention. Lets understand more about global attention mechanisms, for that lets understand about what is attention, local attention before getting to know about global attention mechanism. The summarization approach in (A) and PEGASUS uses local attention and LED uses global attention. So first,about attention, why do we need attention? The fact that a neural network must be able to compress all of the relevant information from a source sentence into a fixed-length vector is a possible challenge with this encoder–decoder approach. Long sentences, especially those that are longer than those in the training corpus, may be challenging for the neural network to cope with. Attention is a method that generates a fuller encoding of the source sequence from which the decoder can construct a context vector. During the prediction of each word in the target sequence, the model can learn which encoded words in the source sequence to pay attention to and to what degree. Now, what is global attention? It is a simplified yet effective attention mechanism which takes the output from the encoder and decoder from the current time step only to calculate attention values. Global attention mechanism is proven to show improved bleu values over local attention/attention. This makes the model capable of producing dynamic length summarizations which helps in handling the MoM case where the meeting may be conducted to any duration and even in case of non-MoM summarizations where context intensity may differ for different input chunks. So it provides a great boost. The arXiv based LED model is also capable of capturing numeric and name based data from the input which was missing in previous approaches and is vital. It was also able to handle noisy data and had a patterned noise exposure in output which was neglected by adding some algorithmic approaches. Other summarization approaches gave noise exposures in outputs without any pattern, which made it difficult to add algorithmic approaches to handle noise since it won't be a generalized solution. The context grap, flow and clarity was better because of evident better bleu adaptation. Further, The ratio of human generated summary compression rate over LED model compression rate is significantly higher. It has some limitations as it shows junk values and repeated sentences at the trail. We were able to remove them with small procedures added after summary processing, and mainly it performed poor on casual talk text inputs. But these aspects are not much impactful on the use case MoM and context summarization. Let's look at one example on the working of LED on text summarization along with word clouds. The input was a text paragraph talking about US governance with a length of just over 900 words. Let's look at the corresponding LED and human generated summary for the same.

LED model:

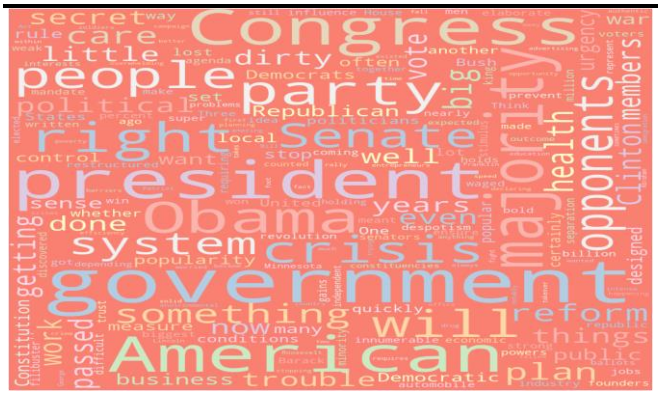
The dirty little secret of American government is that it was designed not to work very well. There are innumerable ways opponents can stop measures from getting passed, even if the president and his party hold a majority in Congress. A minority of 41 senators can "filibuster" a measure and prevent it from coming up for a vote. Republicans lost their majority in Congress in 2006. It often does work. Very well in fact. under the right conditions, barriers fall away and things get done, sometimes with amazing speed and efficiency. What are the right conditions? An overwhelming sense of public urgency. that sense of urgency certainly existed after 9/11, when Congress quickly passed the Patriot Act. getting anything big done in American government requires a sense of crisis. that sense of crisis is why politicians in the U.S. are always declaring crises. they are trying to rally the country to fight a war on something. if the public urgency is not authentic, however, opponents won't have much trouble stopping things from happening.

Human Summarization:

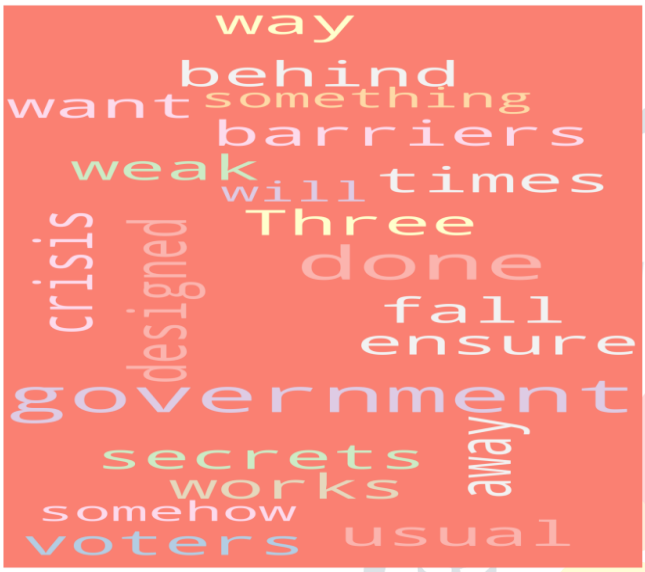
Three secrets behind the way U.S. government works.It was designed to ensure weak government.In times of crisis usual barriers fall away.If voters want something done, it will get done somehow

Now let us look at the word clouds :

- Input



- Human Summary



- LED model summary

We can see that the LED summary actually grabbed a lot of word distribution and context than human summarization in this case. It's similar in many of these cases and we were able to get practical, ideative summaries from the LED model which helps in understanding the target content better.

IV. LITERATURE SURVEY

The paper by (MFA et.al.,) stated that Seq2Seq encoder-decoder architecture is capable of handling context in an extractive way and produces summaries of fixed length for selective domains, but it doesn't imitate how human summarization works. Further, using PEGASUS which uses self-attention mechanism along with Seq2Seq encoder-decoder architecture, was able to provide summaries of fixed threshold lengths and was stable over it. It imitates an abstractive way of summarization, which is close to the human way of summarizing, especially for meeting minutes. But it had its limitations due to the local self attention mechanism which takes context only till an extent, for a segment.It also doesn't provide dynamic summarizations.

The paper by (Quan Wang, et.al., 2018) built a speaker verification system which works on d-vector based approach to speaker diarization. To obtain the state-of-the-art (SOTA) speaker diarization system the combination of LSTM-based d-vector audio embeddings with non-parametric clustering is used.Experiments has been made on both the i-vectors and d-vectors on four clustering algorithms on three standard public datasets and the results stated that d-vector based systems achieve notably lower DER than i-vector based systems.

The patent by (Chong Wang, et.al., 2020) filed a patent for fully supervised speaker diarization, a method consisting of receiving and segmenting the utterance of speech into plurality of segments, the probabilistic generative model is configured and model is trained by including corresponding speaker discriminative embedding and a corresponding speaker label.

The paper by (David Snyder, et.al., 2019) explained an improved version for both speaker recognition and diarization using x-vectors. Adding to that, they implemented a method to withdraw the domain sensitive threshold that is used in the clustering stage of a diarization system. The diarization system is based on x-vectors, PLDA and agglomerative hierarchical clustering (AHC) are used, the performance is measured using the Wild dataset and it is observed that diarization in typically enhanced speaker recognition performance on multi-speaker conversations and preserved powerful performance on single-speaker recordings. They also showed that AHC threshold (controls the number of clusters) can be replaced by different method which accomplish similar performance and the need for an in-domain development set can be eliminated which is used for tuning.

The paper by (David Synder, et.al., 2018) pointed out a method to improve the performance of DNN embeddings for speaker recognition. The easy way of implementation and effective strategy to improve the performance of DNN embeddings for speaker recognition is Data Augmentation. A performance comparison is made on both x-vectors and i-vectors, the results stated that x-vectors outperformed two standard i-vectors baselines on SRE16 Cantonese. The x-vectors got significantly lower error rates than the best on speakers in the Wild after including a large amount of augmented microphone speech. Adding to that they mentioned that x-vectors are great representations for speaker recognition.

The paper by (Tom Ko, et.al., 2017) worked on the comparison of acoustic models that are trained with simulated far-field on a real far-field speech test set. The performance is observed worse in the simulated far-field compared to real Room Impulse Responses (RIRs) and the performance gap is terminated when the point source noises are added. Additionally, they mentioned a substantial improvement can be obtained in the close-talking scenario by combining clean and reverberated training data.

The paper by (Hasim Sak, et.al., 2014) presented a LSTM based RNN architecture model that effectively used to train the acoustic models for the large vocabulary speech recognition. To overcome the issue based on scalability two new architectures were introduced which effectively uses the model parameters compared to the standard LSTM architecture. The two models are introduced,the first as a recurrent projection layer between the LSTM layer and the output layer.The second as non-recurrent projection layer which provides more flexibility by decoupling and to increase the projection layer size without adding more recurrent connections. The effectiveness of the model is significantly higher on a large vocabulary speech recognition task with a large number of output states than the standard LSTM architectures.

The paper by (Jinquing Zhang, et.al., 2020) proposed a pre-trained large Transformer-based encoder-decoder model called PEGASUS which is trained on massive text with a new self-supervised objective. It is used for the exploration of abstractive summarization. Alike Extractive summarization, the key sentences are masked/removed from the input document and one output sequence is generated together from the remaining sentences. The best PEGASUS model is evaluated under 12 downstream summarization tasks and the experiment's results achieved SOTA(State of the art) performance on the 12 tasks measured by ROUGE scores. Additionally, the model gave a surprising performance on the low-resource summarization, by outstripping previous state-of-the-art results. The results are validated finally under human evaluation and it achieved the human level abstractive summarization on various datasets.

The paper by (Iz Beltagy, et.al., 2020) along with Arman Cohan came up with an approach to use global attention mechanism which would scale up linearly on sequence length, this overcomes the limitations of self attention mechanism which won't be able to grasp the context over larger text sequences. This Longformer Encoder Decoder model has been designed to perform better even on larger paragraph inputs, and it also provides dynamic summarizations of variable length, because of the fact that it is able to handle contextual splits by itself.

V. CONCLUSION

In comparison to extractive summary, abstractive summarization produced more intelligible summaries and more stable language phrasing, according to our research review and implementation work analysis. This is because it has been observed that using the same sentence in the input has not been effective in many cases, so abstractive summarization, which focuses on gathering context and rephrasing based on observation, makes more sense towards how humans actually make summarizations, so it correlates with human summarization better than the extractive way. As per this, the Longformer seq2seq Encoder-Decoder transformer model uses global attention mechanism, which even simplifying the mechanism promotes the performance and makes segmentation using speaker diarization or threshold algorithm unnecessary by grabbing the entirety of context flow with linearly extending attention. Noise handling was comparatively simpler in LED because of patternic noise exposure. Moreover, The ratio of human generated summary compression rate over LED model compression rate is significantly higher, which signifies better correlation to human tasks done already, it strengthens the fact that the task done manually currently can be handled by the model too to a surprisingly closer extent in most cases. So the LED model has shown promising results over the other two, its best suited for text summarization than the other two and with further additions and improvements it will be able to satisfy most of the Minutes of Meeting expectation too.

VI. REFERENCES

[1] FM MFA, S P, M G, J J. Automation of Minutes of Meeting(MoM) using Natural Language Processing(NLP), In : 2022 International Conference on Communication, Computing and Internet of Things(IC3IoT). IEEE; 2022. P. 1-6.

[2] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, Ignacio Lopez Moreno, “Speaker Diarization with LSTM”, Google Inc., USA, Carnegie Mellon University, USA , 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), IEEE, Pg 5239-5243, 2018/4/18.

[3] Chong Wang, Aonan Zhang, Quan Wang, Zhenyao Zhu, “Fully supervised speaker diarization”,U.S. Patent 2020/0219517 A1”, Jul. 9, 2020.

[4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, Sanjeev Khudanpur, “Speaker Recognition for Multi-Speaker conversations using X-Vectors”, The Johns Hopkins University,ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, Pg 5796 - 5800, 2019/5/12.

[5] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, Sanjeev Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition”,The Johns Hopkins University, 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, Pg 5329-5333, 2018/4/15.

[6] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, Sanjeev Khudanpur, “A study on data augmentation of reverberant speech for Robust Speech Recognition”, Huawei Noah’s Ark Research Lab, Johns Hopkins University, Microsoft Research, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Pg 5220-5224, 2017/3/5.

[7] Hasim Sak, Andrew Senior, Francoise Beaufays, "Long Short-Term Memory based Recurrent Neural Network architectures for large vocabulary speech recognition", arXiv preprint arXiv:1402.1128, Google, 2014/2/5.

[8] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter Liu, "PEGASUS : Pre-training with Extracted Gap-sentences for Abstractive Summarization",Proceedings of the 37 th International Conference on Machine Learning, PMLR, Pg 11328-11339, 2020/11/21.

[9] Iz Beltagy, Matthew E Peters, Arman Cohan, "Longformer : The Long Document Transformer", arXiv preprint arXiv:2004.05150, 2020/4/10.