Presentation Summarizer: A Full-Fledged NLP Service

Ayush Bhosle[1], Mohd.Raza Deshpande[2], Manas Bokilwar[3], AnasMustafa Dhakwala[4], Prof. Rahul Sonkamble[5]

Dept. of Computer Science & Engineering, MIT School of Engineering, MIT Art, Design and Technology University, Pune - 412201, India

---***---

Abstract - *With the latest advancements in Computational Linguistics, complex Natural Language Processing tasks such as Text Summarization, Language Translation, Text Classification, Question Answering, Grammar Error Correction, etc. have become very feasible. Automatic text summarization on various types of transcripts has proved to be a useful way that best describes the content. However, the traditional methods rely on simpler extractive summarization-based technique. In recent years, Transformers have proved to achieve groundbreaking results, especially for Abstractive summarization task typically known to be involuted.*

Another application where Computational Linguistics has enhanced significantly is Automatic Speech Recognition (ASR), Allowing computers to understand human speech.

Key Words: Sequence to Sequence, Speech Recognition, Grammar Correction, Natural Language Processing.

1. INTRODUCTION

Our Web Service comprises three main components i.e the ASR Module, followed by a Grammar Correction Module and Abstractive Summarizer Module. Users can start transcription (ASR) where speech will continuously be transcribed, and output is displayed in real-time. The Audio Input Stream can be any monologue speech such as a lecture, speech, or conversation.

Once the speech is finished, this transcription is processed by a Grammar Error Correction model since the Kaldi model can only detect words and not the semantic meaning containing punctuations. Hence, reading interpretability is an issue tackled by the GEC model.

In addition, a Named Entity Recognition model will identify entity keywords and highlight the specific part in the text field for visual reading which can be exported by the user to view results later.

Lastly, this corrected transcript is fed to the summarization model to provide a summary.

1.1 Speech Recognition

We have used the VOSK model based on Kaldi ASR. Kaldi ASR is an offline open-source speech recognition toolkit that is utilized for speech-to-text task. It supports 18 language models and dialects, including English, Indian English, German, French, Spanish, Portuguese, Chinese, Russian, Turkish, Vietnamese, Italian, Dutch, Catalan, Arabic, Greek, Farsi, Filipino, and Ukrainian. Kaldi models can have base models (smaller in size) and large models (large size), yet they offer continuous huge vocabulary transcription, low-latency response with streaming API, and changeable vocabulary with speaker identification support.

VOSK model achieves a Word-Error-Rate of ~13 for Indian English.

Fig1: Transcription in Real-Time

1.2 Grammar Error Correction (GEC)

Present Speech Recognition models are only trained to identify spoken terms; hence punctuation marks and prose of the sentence will not be proper. This results in only long words that make reading interpretability difficult.

To tackle this, we incorporate a Grammar Error Correction (GEC) model. This GEC model is a t5-small model trained originally on Wav2Vec2 results mapping incorrect sequence with encoder to a grammatically correct sequence by a decoder.

The Grammar Error Correction approach takes the entire transcript at once and processes the refined text. Since the transcript can be incomplete, it wouldn't contain the proper meaning hence model is applied after the speech utterance has finished, where corrections are made wherever necessary.

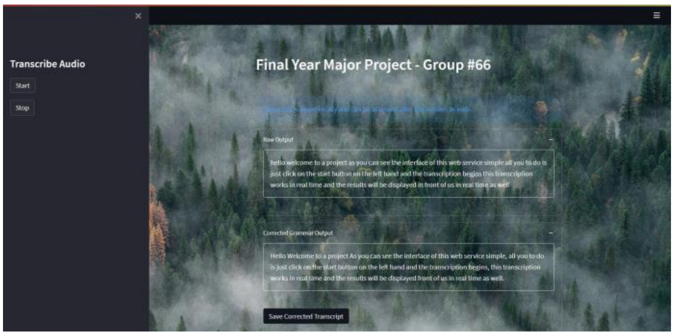

Fig 2: Grammar Error Correction

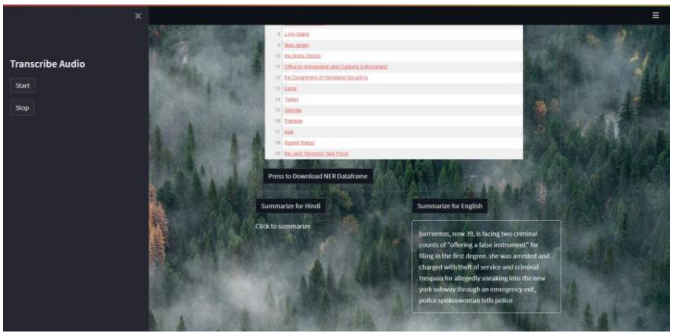

Fig 4: Summary Result

1.3 Text Summarization

For Text Summarization, we have used t5-small model for English language and IndicBART-XLSum for Hindi Language. Both models are based on Abstractive Summarization. The below tables represent the metrics of the pre-trained models with checkpoints uploaded on respective repositories on Hugging Face.

Table -1: IndicBART-XLSum Metrics

Rouge-1	Rouge-2	Rouge-L
0.220394	0.065464	0.198816

Table -2: t5-small Metrics

Rouge-1	Rouge-2	Rouge-L
41.12	19.56	38.35

IndicBART-XLSum is based on extreme summarization, i.e it returns the output that is as small as a couple of lines similar to the headline of a news article.

When Liana Barrientos was 23 years old, she got married in Westchester County, New York. A year later, she got married again in Westchester County, but to a different man and without divorcing her first husband. Only 18 days after that marriage, she got hitched yet again. Then, Barrientos declared "I do" five more times, sometimes only within two weeks of each other. In 2010, she married once more, this time in the Bronx. In an application for a marriage license, she stated it was her "first and only" marriage. Barrientos, now 39, is facing two criminal counts of "offering a false instrument for filing in the first degree," referring to her false statements on the 2010 marriage license application, according to court documents.

Prosecutors said the marriages were part of an immigration scam. On Friday, she pleaded not guilty at State Supreme Court in the Bronx, according to her attorney, Christopher Wright, who declined to comment further. After leaving court, Barrientos was arrested and charged with theft of service and criminal trespass for allegedly sneaking into the New York subway through an emergency exit, said Detective Annette Markowski, a police spokeswoman. In total, Barrientos has been married 10 times, with nine of her marriages occurring between 1999 and 2002.

All occurred either in Westchester County, Long Island, New Jersey or the Bronx. She is believed to still be married to four men, and at one time, she was married to eight men at once, prosecutors say. Prosecutors said the immigration scam involved some of her husbands, who filed for permanent residence status shortly after the marriages. Any divorces happened only after such filings were approved. It was unclear whether any of the men will be prosecuted. The case was referred to the Bronx District Attorney's Office by Immigration and Customs Enforcement and the Department of Homeland Security's Investigation Division. Seven of the men are from so-called "red-flagged" countries, including Egypt, Turkey, Georgia, Pakistan, and Mali. Her eighth husband, Rashid Rajput, was deported in 2006 to his native Pakistan after an investigation by the Joint Terrorism Task Force. If convicted, Barrientos faces up to four years in prison. Her next court appearance is scheduled for May 18.

Fig 3: Sample Input Text

1.4 Named Entity Recognition

Named Entity recognition is the process of identifying keywords that belong to a meaningful category.

Spacy provides a renderer for highlighting keywords present in the text field, Hence, we have used Spacy's Named Entity Recognition model that can scan for entities of various categories such as Location, Money, Date, Cardinal Values, Person Name, Geopolitical Entity, Events and more.

Fig-1: Named Entity Recognition

To enable users to learn more about the entities we also provide a click-able data frame of the unique terms by their type which redirects to a google search.

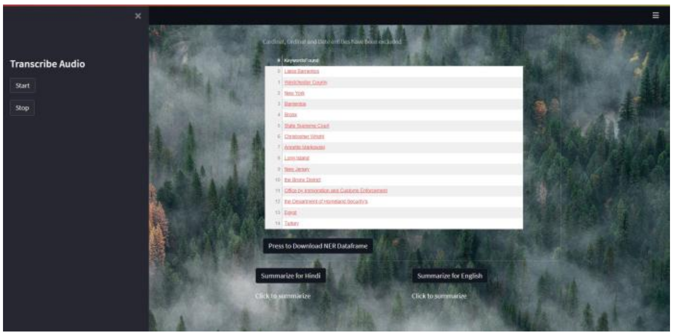

Fig-2: Keywords Dataframe

2. LITERATURE REVIEW

We have researched & tested many State-of-the-Art Transformer & transducer models like Wav2vec2 (Meta), NeMo (NVIDIA), Kaldi based VOSK. One of the main limitations of Transformer Architecture is its capability for real-time transcription, also referred as "Streaming". Since Transformer architecture is based on a self-attention mechanism, these self-attention blocks have a quadratic computational complexity meaning these attention blocks need to look at the complete speech utterance at once making the operation very expensive and hence doesn't achieve streaming.

3. CONCLUSION

Hence, we have successfully completed the development and deployment of the NLP web service on streamlit using various Deep Learning models in pipeline for downstream tasks such as Speech Recognition, Grammar Correction, Text Summarization & Named Entity Recognition.

4. FUTURE WORK

Presently all the models combined add up to a large size which makes storage expensive & deployment complex also requiring GPUs. Hence, we can apply model distillation which converts complex model behavior to a smaller size with respect to parameters. A smaller model will retain only some portion but can be effective for running inferences such as an edge device. Knowledge distillation works best with Natural Language Processing models.

For speech recognition, for use cases that deal with specific vocabulary, speech samples can be trained to learn the vocabulary. Kaldi also supports Speaker Diarization, meaning identifying the speech spoken by the speaker expanding the scope to even wider use cases.

REFERENCES

[1] Alexei Baevski, H. Z. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv: 2006.11477 [cs.CL].

[2] Oleksii Kuchaiev, Jason Li, Huyen Ngu-yen, et al(2019), NeMo: a toolkit for building AI applications using Neural Modules, arXiv:1909.09577 [cs.LG].

[3] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, et al. (2020), End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures, arXiv:1911.08460 [cs.CL].

[4] Satoshi Kaki, Eiichiro Sumita, and Hitoshi Iida (1998). A Method for Correcting Errors in Speech Recognition Using the Statistical Features of Character Co-occurrence. In COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics.

[5] Wang, Xiuhua & Zhong, Weixuan. (2022). Research and Implementation of English Grammar Check and Error Correction Based on Deep Learning. Scientific Programming. 2022. 1-10. 10.1155/2022/4082082.

[6] Zhu, J., Uren, V., Motta, E. (2005). ESpotter: Adaptive Named Entity Recognition for Web Browsing. In: Althoff, KD., Dengel, A., Bergmann, R., Nick, M., Roth-Berghofer, T. (eds) Professional Knowledge Management. WM 2005. Lecture Notes in Computer Science, vol 3782. Springer, Berlin, Heidelberg. doi:10.1007/11590019_59

[7] Lizarralde I, Mateos C, & Rodriguez JM, Zunino A (2019). Exploiting named entity recognition for improving syntactic-based web service discovery. Journal of Information Science;45(3):398-415. doi:10.1177/0165551518793321

[8] Imran Sheikh, Emmanuel Vincent, & Irina Illina. On semi-supervised LF-MMI training of acoustic models with limited data. INTERSPEECH 2020, Oct 2020, Shanghai, China. ffhal-02907924f.

[9] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, et al (2016). "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in Interspeech, pp. 2751–2755.

[10] A. Carmantini, P. Bell, and S. Renals, (2019) "Untranscribed web audio for low resource speech recognition," in Interspeech, pp. 226–230.