Hindi to English Neural Machine Translator and Extractive Summarizer

Samyak Maurya
Computer Science
PES University
Bangalore, India
samyakm51@gmail.com

Vaibhav G
Computer Science
PES University
Bangalore, India
vaibhavg.01@gmail.com

Navya Agrawal
Computer Science
PES University
Bangalore, India
navyaagrawal01@gmail.com

***Abstract*—This project aims to build a News Translator and Summarizer using Neural Machine Translations which uses a sequence to sequence model which works on the encoder decoder architecture to translate news articles from Hindi to English.TF-IDF is used to summarize the translated text which makes it convenient for the reader to skim through articles.Further we have used gTTs a python library to convert the summarized text into Speech. Results obtained by the extractive summarizer were up to the mark and that of the machine translator can use more improvement.**

I. Introduction

In the post-modern word, translation has become so relevant that people often visualize it as a socio-cultural bridge between communities and countries.Translation has helped unite India together as a nation throughout her history. Ideas and concepts like 'Indian literature', 'Indian culture', 'Indian philosophy' and 'Indian knowledge systems' would have been impossible in the absence of translations.In a multilingual country like India, translation helps people to read and understand news from various parts of the country and different part of the world as well.Although accessing news is easy, it can be a challenge for people to translate it into their suitable language.This project would like to solve this issue using Neural Machine Translation to provide a platform for people to understand news that could have not been comprehensible before.

By further adding the feature of Summarization, the reader can save time and can also decide whether to go through an entire document or not.The main goal of newspaper article summary is, the readers to walk away with knowledge on what the newspaper article is all about without the need to read the entire article.Text to speech translation will help with extracting insights without the user specifically reading them. This project focuses on building a model that can help with such challenges.

For translating news articles we use Neural machine translation (NMT) which is an approach of machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model. The translator uses the sequence to sequence(Seq2Seq) model that aims to map a fixed-length input with a fixed-length output where the length of the input and output may differ. This has been done using the encoder-decoder architecture which trains a single end-to-end model directly on source and target sentences. For Text Summarization an Extractive Summarization technique is applied that uses Term Frequency-Inverse Document Frequency(TF-IDF), a numeric measure which is used to score the importance of a word in a document based on how often it appears in that document and a given collection of documents. The text to speech was done using a simple python library which helps to convert the summarized news articles into speech.

The Data set used for this project was obtained from anki.net. It contains datasets for several lan-

guages which map to English. For this project we have used the English-Hindi dataset which consists of 2872 translated sentence pairs from English to Hindi, each language separated by a tab and each pair separated by a newline character, which has about 3556 English words and 3211 Hindi words. The PyTorch library is used for translation as its easier to integrate with python,

II. Related Work

A. Pilault, Jonathan, et al. "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, https://doi.org/10.18653/v1/2020.emnlp-main.748.

This paper aims to produce abstractive summaries of long documents(journals) that exceeed several thousand words using neural abstractive summarization. They performed simple abstractive step before generating summary. The objective of this extractive step was to condition the transformer language model on relevant information before being tasked with generating summary. This step uses a 2 layered encoder using bidirectional LSTM's to generate sentence embeddings and document representations. The decoder uses attention mechanism which produces an output which is a function of attention weights obtained by previous timestamps and current decoder input and previous hidden state. The journal to be summarized is then divided into 4 parts, the introduction and the extractive summary that acts as a proxy for generating enough information for summary and the rest of the paper that acts as the domain. This data was then fed to the transformer learning model which produced a confidence of 95 percent on 4 summarization datasets. This model was compared with previously existing work by means of Rouge scores and showed considerable improvement. The main draawback of this model is abstractive summaries generated by transformers can generate imaginary content and extensive usage of hardware like GPU's are required to train them. Also abstractive summary for selecting tokens can tend to repeat throughout the document summarizing only a certain portion of it.

B. Gehlot, Akanksha, et al. "Hindi to English Transfer Based Machine Translation System." Https://Arxiv.Org/, vol. 5, no. 19, 2015, https://doi.org/10.48550/arXiv.1507.02012.

This paper aims to translate Hindi Text to English using Transfer and Rule based approaches. This project used CYK algorithm to translate hindi text to english using Context Free Grammar(CFG) in Chomsky Normal Form(CNF). The hindi sentences were pos tagged and CYK algorithm was used to find equivalent english translation. Few rules like transliteration and morphological analysis were used for tokens and rules for dealing with interrogative sentences, replicative nouns and synonyms were introduced by the authors. The project fails to incorporate rules for dealing with idioms and complex compound sentences and have mentioned it as future scope. The project fails to address metonymy. Apart from these drawbacks, we cannot find CFG's for different languages easily as a lot of time is spent on making them

C. Hayashi, Tomoki, and Shinji Watanabe. "DiscreTalk: Text-to-Speech as a Machine Translation Problem.", Human Dataware Lab. Co. Ltd., Japan Nagoya University, Japan Johns Hopkins University, USA, 12 May 2020, https://arxiv.org/pdf/2005.05525.pdf.

This research proposes a new strategy for an end-to-end text to voice (E2E-TTS) conversion model based on a mix of a non autoregressive vector quantized variational autoencoder (VQ-VAE) model and an autoregressive transformer NMT model. The Transformer-NMT model is trained to predict the discrete symbol sequence from a given input text after the VQ-VAE model learns a mapping function from a speech waveform into a sequence of discrete symbols. An encoder, codebook, and decoder make up the VQ-VAE model. The encoder turns the down sampled N-length sequence of vectors to a discrete symbol sequence, and the decoder maps the non-liner function to reconstruct the input waveform from quantized vectors. The beam search and subword units techniques developed in NMT and automatic speech recognition (ASR) were used in this paper. The training phase and the synthesis phase are the two stages of the method. The VQ-VAE model is trained utilising speech waveforms

from the corpus during the training stage. Text is used to train the NMT model, and subword sequences are used as the target. The transformer NMT model uses beam search to estimate subword sequence from a given text, after which the sentence piece model estimates discrete symbols, which are then fed to the decoder, which converts into speech waveform. There is a trade-off between DSF and speech articulation, and the model does not extend to the multi-speaker framework.

D. Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, Eunah Cho, Matthias Sperber, Alexander Waibel. "The Karlsruhe Institute of Technology Systems for the News Translation Task in WMT 2017." Proceedings of the Second Conference on Machine Translation, 2017.

This study discusses news translation from three languages: English to German, German to English, and English to Latvian. The neural machine translation model based on encoder-decoder is utilised here. Open NMT, an eTorch-based toolset, is used in the project. This toolkit, which contains two LSTM layers of 1,024 units, is used to train each model. The gradients are scaled at 5, and Adam is employed experimentally with a high learning rate of 0.001, which is then reduced to 0.0005 until the model's perplexity does not decline any further. Every epoch, checkpoints are saved, and the toolkits have been expanded with new capabilities, such as the Context Gate for attention modelling and the use of coverage information during learning to translate from one language to another. The current NMT system's biggest flaw is its restricted vocabulary, which makes it difficult to translate unusual terms. The BLEU ratings for translating from German to English and English to Latvian are respectively 38.39 and 24.11.

E. Zhu, Junnan, Yu Zhou, Jiajun Zhang, Chengqing Zong . "Attend, Translate and Summarize: An Efficient Method for ..." Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization, 2020, https://aclanthology.org/2020.acl-main.121.pdf.

This project uses cross-lingual summarization to summarise a document written in one language into another. The authors focused on the original terms first, then gathered translation candidates and included them into the final summary. To attend to some words and obtain translation candidates from a probabilistic bilingual lexicon, the first encoder-decoder attention distribution was used. Then a translating probability p(trans) was calculated, which balanced the probability of generating words from the neural distribution with that of selecting words from the translation candidates of the source text. The final distribution was obtained by the weighted sum (weighed by p(trans) of the neural distribution PN and the translation distribution PT). Few limitations were found such as this method doesn't incorporate into the multi-task method. Only covers summarization and we require much more functionalities.

F. Vijayakumar, K. P. Vijayakumar, Hemant Singh, Animesh Mohanty. "Real Time Speech to Text Text to Speech Converter with Automatic Text Summarizer Using Natural Language Generation and Abstract Meaning Representation." International Journal of Engineering and Advanced Technology, vol. 9, no. 4, 2020, pp. 2361–2365., https://doi.org/10.35940/ijeat.d7911.049420.

This paper examines real-time Speech-to-Text and Text-to-Speech conversion, as well as automatic text summarization, using Natural Language Generation and abstract meaning representation techniques. This approach translates real-time voice to text, then to a summary using Natural Language Grammar (NLG) and Abstract Meaning Representation (AMR) graphs, before converting the summary back to speech. Deep Speech 2 and AMR graphs are two primary techniques used in the proposed paper. When the speech recognition model was run on the Central Processing Unit (CPU), the speedup was 4x, and when the deep learning techniques were executed on the Graphics Processing Unit (GPU), the speedup was 21x. The following are some of the limitations: AMR parsers and generators of higher quality are still required. Better-accurate parsers and generators can considerably boost the model's efficiency. There is a requirement to summarise the spoken text, however there are no acceptable datasets that could increase the output significantly.

III. SCOPE OF WORK

India is a multilingual country with over 22 official languages, but several other unofficial languages changing from region to region. Hence a news article of interest will not be understood by the majority of Indians. Hindi and english are recognized as the official language of India but a considerable part of the country cannot read hindi. This is where the motivation of our Neural Machine translator comes from.
Several articles in newspapers or over the internet may not be of use to viewers who want to go through a specific news article. Hence our summarizer can prove beneficial saving a lot of unnecessary time being spent on searching for the right news article.

IV. PROPOSED METHODOLOGY

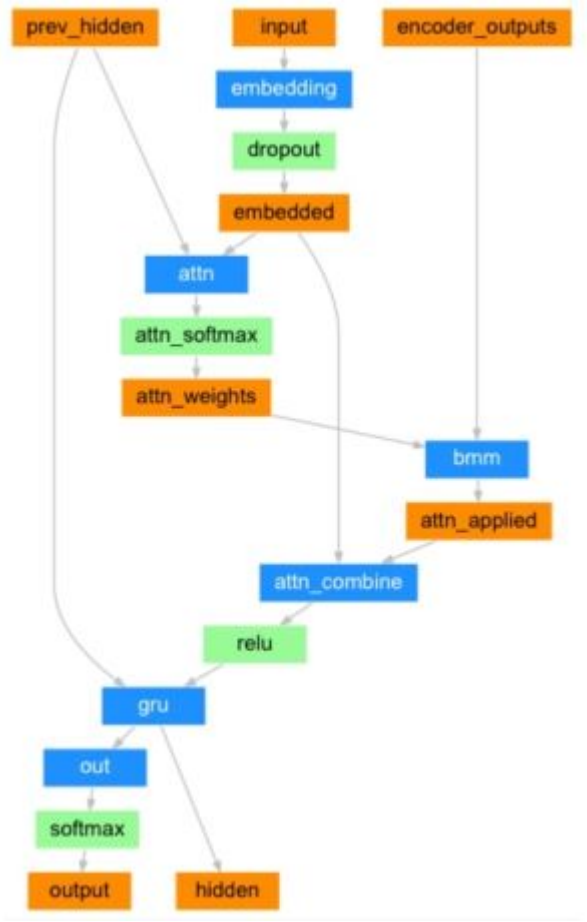

Fig. 1. Attention layer: focuses on extracting key information. Uses drop-out regularizer to prevent over-fitting

A. NMT Architecture

- Seq2Seq Model, A Recurrent Neural Network (RNN) is a network that runs on a sequence and feeds its own output into subsequent steps.
- Unlike sequence prediction with a single RNN, where each input corresponds to an output, the seq2seq model is unconstrained by sequence length or order, making it ideal for translation between two languages. The encoder in a seq2seq model creates a single single vector that, encodes the ”meaning” of the input sequence into a single vector — a single point in an N-dimensional space of sentences.
- A seq2seq network’s encoder is an RNN that outputs a value for each word in the input phrase. The encoder outputs a vector and a hidden state for each input word, and the hidden state is used for the following input word.
- If the encoder and decoder just exchange the context vector, that single vector is responsible for encoding the entire phrase.
- For each step of the decoder’s own outputs, the decoder network can ”focus” on a different component of the encoder’s outputs. We begin by determining a set of attention weights. As indicated in figure [2] this will be multiplied with the encoder output vectors to generate a weighted combination.
- The result should provide information about that portion of the input sequence, assisting the decoder in selecting the appropriate output words. The attention weights are calculated using the decoder’s input and hidden state as inputs in another feed-forward layer called attn. Shorter sentences will only employ the first few (weights as shown in figure [1]), whereas longer phrases will use all of them.

$$\alpha_{ts} = \frac{\exp(score(h_t, h'_s))}{\sum \exp(score(h_t, h'_{s'}))} \quad (1)$$

$$c_t = \sum \alpha_{ts}.h'_s \quad (2)$$

$$a_t = f(c_t, h_t) = \tanh W_c[c_t; h_t] \quad (3)$$

$$score(h_t, h'_s) = v_a^\top \tanh W_1 h_t + W_2 h'_s \quad (4)$$

(1)-Attention weights
(2)-context vector

(3)-Attention vector
(4)-Bahdanau's additive style

- The attention decoder also optionally uses the teacher forcing algorithm. This algorithm allows the next input of the decoder to be a target token from the data set rather than the output of the previous timestamp. The final output of this model is the english text corresponding to the hindi embeddings.

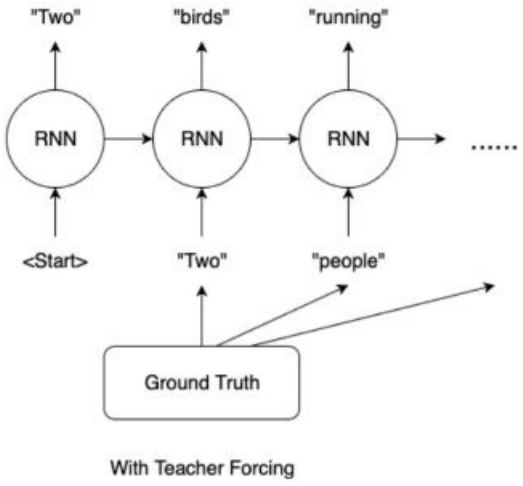

Fig. 2. Teacher Forcing method: Uses Actual output obtained from ground truth as next input for every timestamp

B. Text Summarizer

- The TF-IDF is a numerical statistic that indicates how essential a word is to a document in a corpus or collection. The TF-IDF score rises in proportion to the number of times a word appears in the document, but this is counterbalanced by the word's frequency in the corpus, which helps to regulate the fact that some words are more common than others.
- The raw frequency of a term in a document is referred to as the frequency term. Inverse document frequency is a metric for determining if a phrase is common or uncommon across all documents, which is calculated by dividing the total number of documents by the number of documents containing the term.
- The TF-IDF value of each noun and verb can then be determined using the preprocessed list of words. The TF-IDF equation is shown below.

$$TF = \frac{Totalappearanceofawordindocument}{Totalwordsinadocument}$$

$$\text{IDF} = \log \frac{AllDocumentNumber}{DocumentFrequency}$$

$$TF - IDF = TF \times IDF$$

C. Text to speech Translator

- This project uses Googles open source python library called gTTs(Google text to speech) to convert text to audio in a mp3 format. Work is being done to use the vector quantized variational autoencoder(VQ-VAE) model and an autoregressive transformer NMT model which uses advance decoding techniquessuch as Beam-search, Shallow fusion with LM and subword units commonly used in NMTs.

V. Results

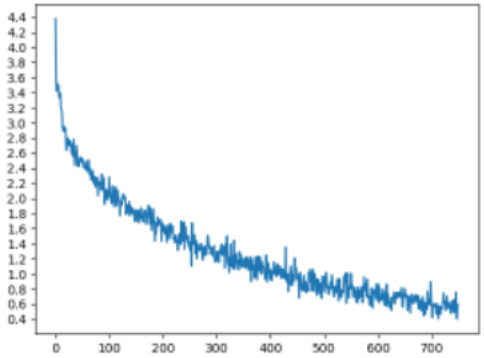

Fig. 3. Plot of Loss achieved after 75000 iterations

A. NMT

We implement out Neural Machine Translator using the PyTorch library as its more convenient for python programming.
We ran our model for 75000 and 40000 iterations noticing significant differences with both.
On running with 75000 iterations we found that the model was slightly over-fitting the training data and as seen in Fig.4 the loss reduces drastically and comes down to 0.085 from 4.4 which shows that the model performs with a decent accuracy.
On running 40000 iterations the model was slightly under-fitting the training data and was also

performing poorly on the testing data.

```
> तुम्हारे घर का फ़ोन नम्बर क्या है?
= What's your home phone number?
< What the call in your age? <EOS>

> मेरे दोस्तों ने मेरा जन्मदिन मनाया।
= My friends celebrated my birthday.
< My friends celebrated my birthday. <EOS>

> मुझे उसकी चिट्ठी का जवाब देना है।
= I have to answer his letter.
< I have to answer his letter. <EOS>
```

Fig. 4. Results observed with 40000 iterations

```
> वह नीचे झुकी।
= She bent down.
< She bent down. <EOS>

> वह हर सुबह अपने कुत्ते को सैर पर ले जाता है।
= He walks his dog every morning.
< He walks his dog every morning. <EOS>

> मेरा पेट भर गया है।
= I'm full.
< I'm full. <EOS>
```

Fig. 5. Resluts observed with 75000 iterations

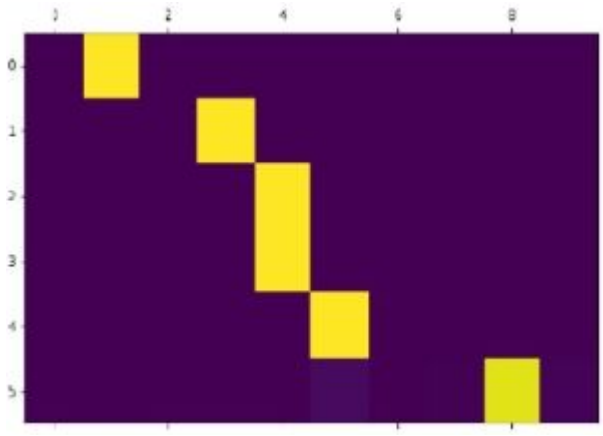

Fig. 6. Visualizing attention

As observed model 2 clearly outperformed model 1

B. Text Summarizer

We obtained data for summarization via a python library called newspaper3k. Using Tf-Idf algorithm and taking top n sentences,we were able to accurately get the summary of an entire news article/

VI. Conclusion

The main objective of our project was to knit the linguistic diversity in India. Citizens from different states who were separated from different language boundaries can now read news articles in other languages as well, This can also be extended for International news where one could understand news articles written in different foreign languages. This project aims to translates the Hindi news articles to English and further summarizes them with a fairly good accuracy. The data set used for training the translator was English-Hindi dataset from anki.net. The translator performed fairly well which gave an error rate of 0.085 when trained for 75000 iterations.

A. Future scope

- The data set consisted of 2872 translated pairs which was not adequate to train our model to provide the best results, use of larger data sets required high performance computation like GPU's which were not used.
- To improve the performance of the E2E TTS, we further aim to use the combination of a non autoregressive vector quantized variational autoencoder(VQ-VAE) model and an autoregressive transformer NMT model which performs better than the conventional method.
- The use of dropout function in the attention decoder helped to reduce overfitting by a very small margin hence we need to come up with better techniques to reduce this issue.

References

[1] Pilault, Jonathan, et al. "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, https://doi.org/10.18653/v1/2020.emnlp-main.748.

[2] Gehlot, Akanksha, et al. "Hindi to English Transfer Based Machine Translation System." Https://Arxiv.Org/, vol. 5, no. 19, 2015, https://doi.org/10.48550/arXiv.1507.02012.

[3] Hayashi, Tomoki, and Shinji Watanabe. “DiscreTalk: Text-to-Speech as a Machine Translation Problem.”, Human Dataware Lab. Co. Ltd., Japan Nagoya University, Japan Johns Hopkins University, USA, 12 May 2020, https://arxiv.org/pdf/2005.05525.pdf.

[4] Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, Eunah Cho, Matthias Sperber, Alexander Waibel. “The Karlsruhe Institute of Technology Systems for the News Translation Task in WMT 2017.” Proceedings of the Second Conference on Machine Translation, 2017.

[5] Zhu, Junnan, Yu Zhou, Jiajun Zhang, Chengqing Zong . “Attend, Translate and Summarize: An Efficient Method for ...” Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization, 2020, https://aclanthology.org/2020.acl-main.121.pdf.

[6] Vijayakumar, K. P. Vijayakumar, Hemant Singh, Animesh Mohanty. “Real Time Speech to Text Text to Speech Converter with Automatic Text Summarizer Using Natural Language Generation and Abstract Meaning Representation.” International Journal of Engineering and Advanced Technology, vol. 9, no. 4, 2020, pp. 2361–2365., https://doi.org/10.35940/ijeat.d7911.049420.