# CS626 Course Project: Cross-Lingual Summarization

Inderjeet Jayakumar Nair, 170020013
Saurabh Jayesh Parekh, 170100016
Manas Jain, 170040068

December 2020

## 1 Abstract

We implement a transformer based architecture for cross-lingual summarization using shared encoder-decoder architecture. The traditional approaches involves sequential application of machine translation followed by machine summarization or machine summarization followed by machine translation. However, such a method would result in tremendous usage of memory and would result in error propagation. We also create novel dataset for the purpose of cross lingual summarization from english to hindi. We also contribute by providing the code for a novel state of the art architecture for cross-lingual summarization whose implementation is currently unavailable.

## 2 Introduction

Imagine yourself in a foreign train station rife with a variety of instruction boards. Suppose you need to find the instructions for booking a ticket and all the instructions are in a Foreign language. The different types of instructions may include safety instructions, instructions to use a particular equipment, ticket pre-booking instructions, etc. Directly translating the content would produce a large chunk of text which would not be user friendly. We aim to alleviate this issue by proposing a model that summarizes the text from one language to another by using a single transformer. The use of only a single transformer makes our model very light, memory efficient, fast and easily deployable in mobile devices.

As already stated, the current techniques involves sequential usage of machine translation followed by machine summarization or machine summarization followed by machine translation. This results in heavy requirement for the computation resources and often results in error propagation. Thus there is a need for faster models which uses significantly lesser memory and offers faster computation. This implies using shared architecture for encoder-decoder for performing cross-lingual summarization. There are two variants in applying this method: RNN based sequential model and Transformer models with parallel computability. We experiment with both the models in our project and confirm that Transformer in general provides better results and theoretically, due to high level of parallelization, Transformer are significantly faster.

## 3 Problem Statement

Consider two Languages: Source Language($L_S$) and Target Language($L_T$). We are given with a sentence $S_S$ as input from $L_S$ which may be assumed to be grammatically valid. We define our problem statement as proposing a model that generates a sentence $S_T$(grammatically valid) as output in $L_T$ such that

1. $S_T$ semantically captures the most relevant / important section of $S_S$
2. The number of tokens in $S_T$ is less than the number of tokens in a valid machine translated output of $S_S$ from $L_S$ to LT

If $L_S = L_T$, the task becomes automatic summarization. We want to perform this task using shared encoder-decoder architecture.

# 4 Dataset Information

## 4.1 HindiEnCorp 2.0

1. Hindi-English Parallel Corpora with **127,607** sentences
2. Pre-trained models with this dataset to learn the language model as the pretrained models for hindi were difficult to find
3. **Hypothesis behind using the dataset:** If the language model is learnt in the encoder and decoder, cross lingual summarization dataset can just fine tune the attention weights in order to capture important segments, thereafter the encoder with learned language model can just obtain valid output corresponding to the important segments.

## 4.2 Daily News Dataset

1. News-headlines dataset in English with **90,000** training sentences and **4,515** test sentences
2. Machine translated headlines to Hindi by using free **Google Translate API**
3. The machine translation was done in a batch of 4,500 sentences since the API was unstable

# 5 Dataset Examples

## 5.1 HindiEnCorp 2.0

1. Humans destroyed the commons that they depended on : मानवों ने उन ही साझे संसाधनों को नष्ट किया जिन पर वो आधारित थे।
2. Premchand is the author of the Hindi literature era : प्रेमचंद हिन्दी साहित्य के युग प्रवर्तक हैं |

## 5.2 Daily News Cross lingual summarized

1. The police on Saturday arrested three persons from a gang of five alleged criminals in Greater Noida, after a shootout in which one of them was injured and two managed to flee. The police had acted on a tip-off and discovered later that the gang from Bulandshahr, was involved in a robbery in Noida's Kasna this month : ग्रेटर नोएडा में पुलिस के साथ गोलीबारी के बाद हुए 3 लुटेरे
2. Finance Minister Arun Jaitley on Friday said that the financial year may soon begin from January in sync with the calendar year. ""The matter of changing financial year is under consideration of the government,"" he told the Lok Sabha. In May, Madhya Pradesh became India's first state to shift its financial year format to January-December from the present April-March cycle.: जनवरी से शुरू हो सकता है वित्तीय वर्ष: जेटली

# 6 Approaches

1. Sequence-2-Sequence GRU with attention mechanism [1]
2. Transformer Architecture [2]
3. Multi-Tasking Transformer Architecture [3]

In the first and second approach, we follow the following training strategy:

- Pretraining on Machine Translation Corpus
- Fine tunning for cross lingual summarization

We wish to explain the hypothesis behind following such strategy. Training the model to perform machine translation would make the encoder learn the langauge model for English and the decoder to learn the language model for Hindi. Thus, having learned the language models for the source and target languages, we hypothesize that cross lingual summarization can be enabled by fine tuning the attention weights, wherein, the decoder can focus on important aspects of the sentences and would produce meaning sentence in the target language as it has learned the corresponding language model.

The third approach is also a Transformer-based approach with 2 decoders and 1 encoder. This results in the formulation to 2 encoder-decoder modules with the same encoder. One of the pairs would focus towards learning machine translation task and the other pair towards cross-lingual summarization. We choose these tasks for two reasons:

- The task of machine translation can be regarded as cross-lingual summarization with a compression ratio of 1:1. Thus learning both the tasks simultaneously would benefit both the tasks with an added benefit as they make the encoder more specialized in understanding the source language. With such a superior encoder, the model may perform better.
- This allows us to use the machine translation dataset and cross-lingual summarization dataset simultaneously thus speeding up the training process.

### 6.1 Seq-2-seq GRU with attention

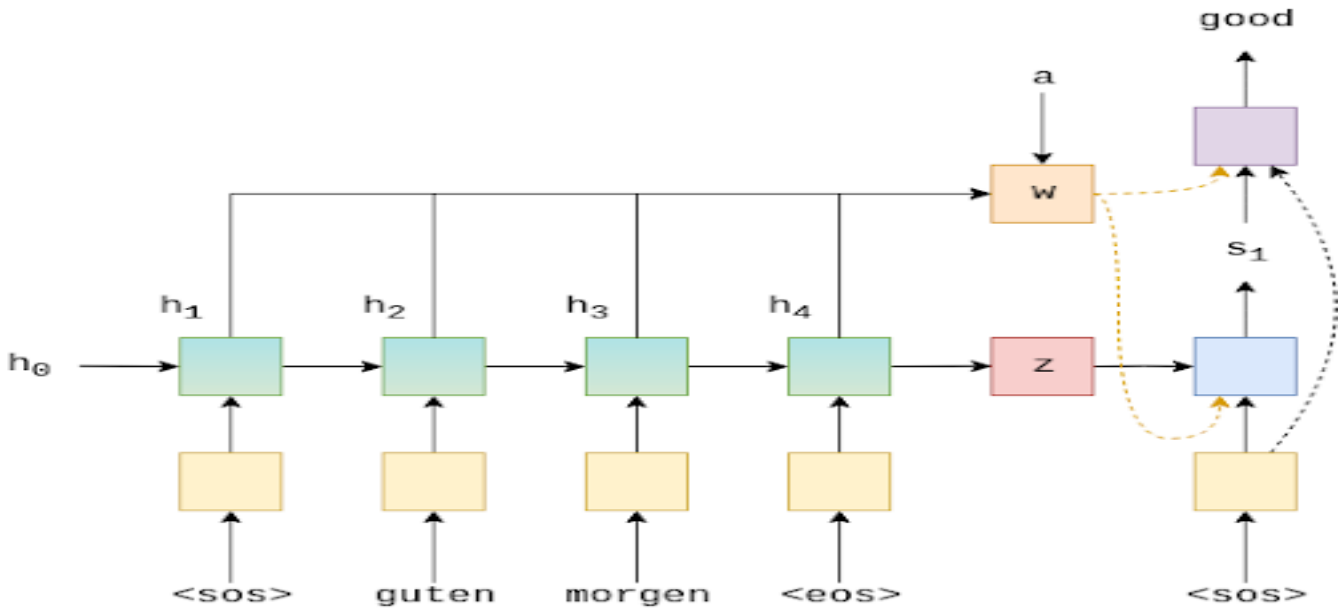


Figure 1: Seq-2-seq GRU with attention architecture

#### 6.1.1 Encoder

We use a single layer GRU, however we now use a bidirectional RNN. With a bidirectional RNN, we have two RNNs in each layer. A forward RNN going over the embedded sentence from left to right (shown below in green), and a backward RNN going over the embedded sentence from right to left (teal).

#### 6.1.2 Attention

This will take in the previous hidden state of the decoder and all of the stacked forward and backward hidden states from the encoder, The layer will output an attention vector, , that is the length of the source sentence, each element is between 0 and 1 and the entire vector sums to 1. Intuitively, this layer takes what we have decoded so far, , and all of what we have encoded to produce a vector that represents which words in the source sentence we should pay the most attention to in order to correctly predict the next word to decode

### 6.2 Transformer Architecture

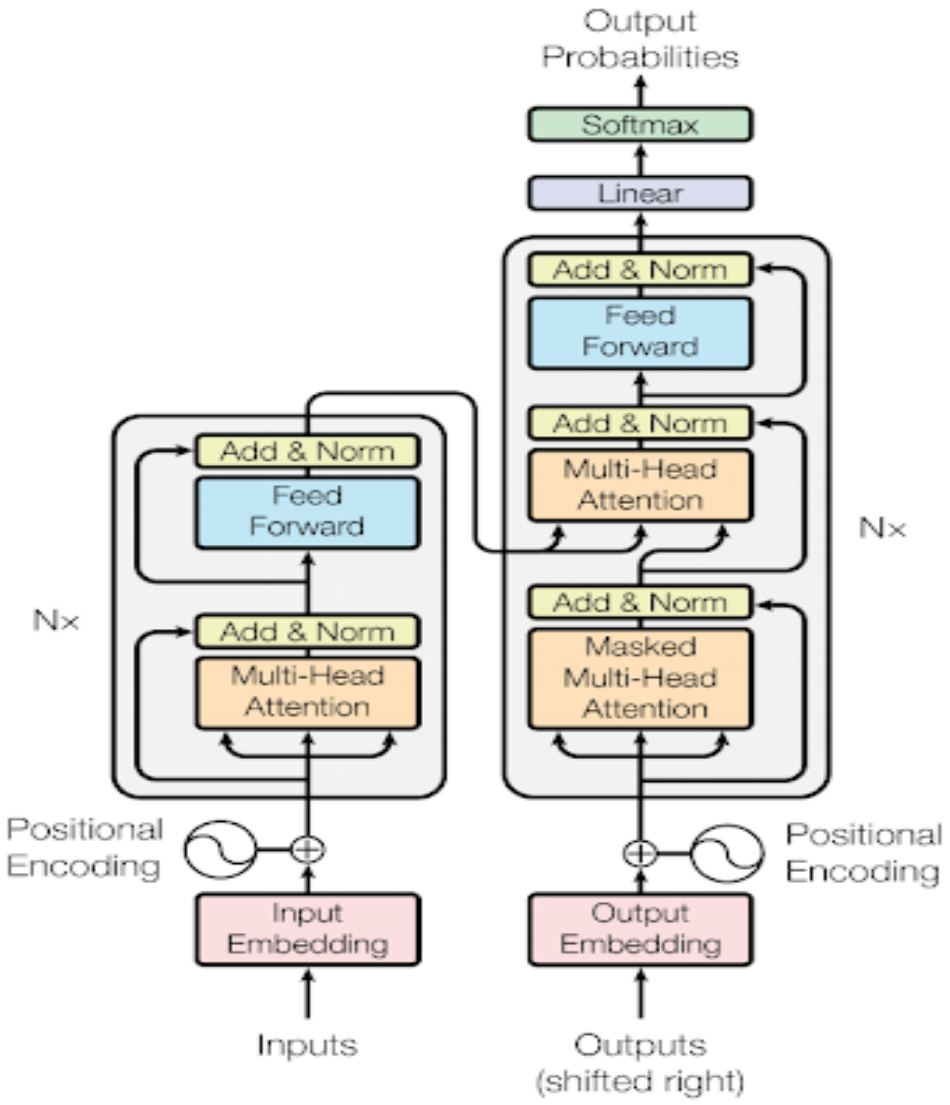


Figure 2: Transformer Architecture

1. Pretrained on HindiEnCorp 2.0
2. Trained on Daily News
3. Total Pretraining time = 3 days
4. Total Training time = 3 hrs

### 6.3 Multi-Tasking Transformer Architecture

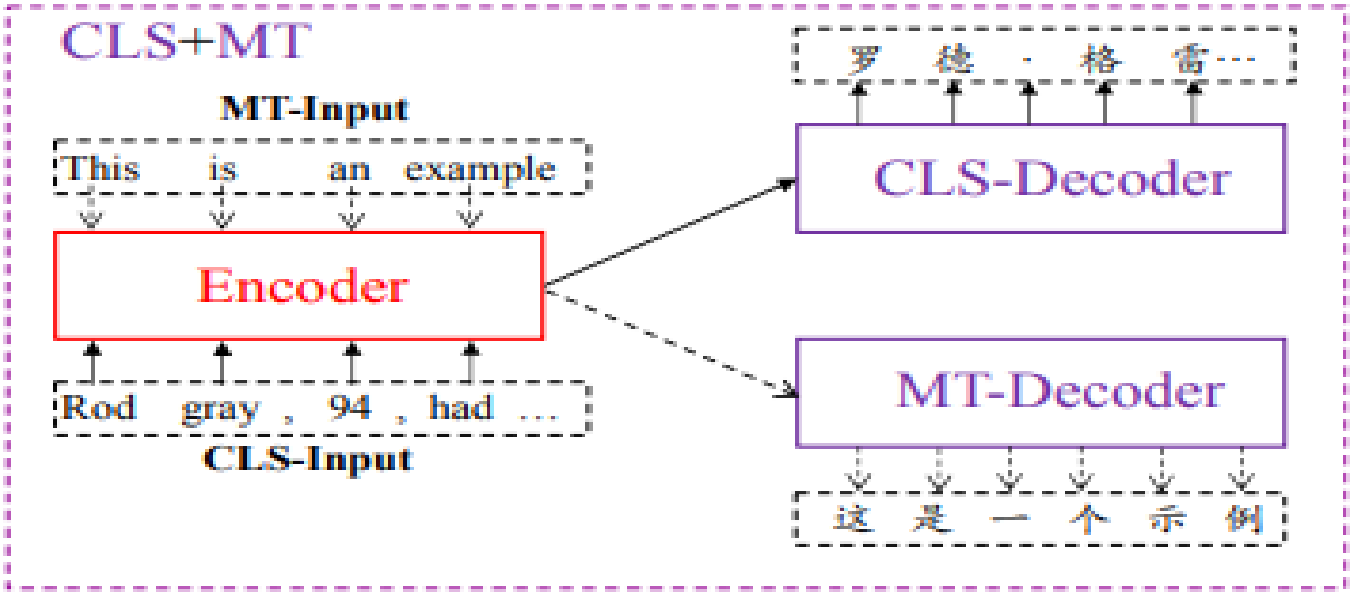


Figure 3: Multi-Tasking Transformer Architecture

In this approach, a multi-task loss function objective is set up to train cross-lingual summarizer. The transformer consists of one encoder and two decoders for the task of machine translation and cross-lingual summarization. Since MT can be regarded as a special case of CLS with compression ratio 1:1. Jointly training the model makes the encoder highly specialized in language understanding.

1. Implemented the model from scratch in Pytorch
2. Jointly trained on HindiEnCorp 2.0 MT dataset and Daily News CLS dataset
3. The model is still under training due to the requirement of high computational power for such specialized architecture
4. Achieved 0.3406 BLEU score when just trained for 3 epoch

## 7 Some examples from the Trained Transformer Model

1. **Input:** Four labourers on Monday were reportedly injured after a tree branch fell on them at Dombivli station road in Mumbai. They were admitted to hospital with injuries and were later declared out of danger. Reportedly, tree fall cases are on rise in Kalyan-Dombivli. Last year fewer cases were reported. We have been getting complaints of tree falls daily.
   **Output:**मुंबई में पेड़ की शाखा गिरने से 4 मजदूर घायल हो गए
2. **Input:** The Bombay High Court on Monday summoned the Maharashtra Women and Child Development Secretary after 42 children went missing over the last three years from a Mumbai remand home. The court criticised the Maharashtra government for lack of 'proactive action' in the matter. The Bombay High Court is hearing a PIL on the allegations of corruption in the remand home.
   **Output:** 42 बच्चों के घर से बाहर निकलने के बाद बॉम्बे <unk> ने सचिव को बुलाया
3. **Input:** As many as 76 passengers were rescued from cable cars suspended over a river in German city Cologne after a gondola crashed into a support pillar on Sunday. Passengers were left stranded, and children were seen clinging to parents while dangling as many as 40 metres above the river. The fire department lowered them to safety from the cable cars
   **Output:** केबल कार <unk> से उतरने के बाद 76 यात्री निलंबित

4. **Input:** An 11-year-old tribal boy allegedly committed suicide on Tuesday by hanging himself near his school, after he was caught stealing ?30 from his classmate in Maharashtra's Mokhada. The boy was reportedly ashamed of his act and had tried to force a classmate to commit suicide with him, but he refused. Police said the boy has a history of criminal activities.
**Output:** 19 – वर्षीय आदिवासी लड़के ने <unk> पकड़ा , 30 से 30 पकड़े जाने के बाद आत्म हत्या कर ली

# 8 Challenges

1. **Dataset Generation:** Unable to find dataset for English-Hindi CLS
2. **Limited Computation power:**
   (a) This made us use lesser number of words in the vocabulary as the number of parameters increases with the size of the vocabulary. Thus in many of the previous examples we obtained <UNK> token corresponding to unknown word.
   (b) Huge training time is required for seq2seq and Multi task Transformer and thus they are being trained
   (c) We are unable to use the CFILT parallel Hindi English Corpus for pretraining due to large memory requirements
3. **Google's Rate Limiter:** The free API makes several HTTP requests to the Google's machine translation server to retrieve the machine translated output. As a result several times, our IP was blocked from using Google Services
4. **Unstable Session of Colab:** When the code is executing for a large period, the session collapses to limit the usage. Colab prevents us from over using their GPU resources by disabling our access to GPU

# References


[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks." In: Advances in neural information processing systems. 2014, pp. 3104–3112.

[2] Ashish Vaswani et al. "Attention is all you need." In: Advances in neural information processing systems 30 (2017), pp. 5998–6008.

[3] Junnan Zhu et al. "NCLS: Neural cross-lingual summarization." In: arXiv preprint arXiv:1909.00156 (2019).