Code-Switched Hindi-English Machine Translation

Manas Bundele (mmb180005)
Randeep Ahlawat (rsa190000)
Siddhartha Sahai (sxs180170)

Task Definition

Hindi-English : “Main kal movie dekhne jaa rahi thi and raaste me I met Sam.”

Gloss : I yesterday [movie] to-see go Continuous-marker was [and] way in [I met Sam].

English Translation : I was going to a movie yesterday and I met Sam on the way.

Machine translation of code-switched Hindi-English sentences to English

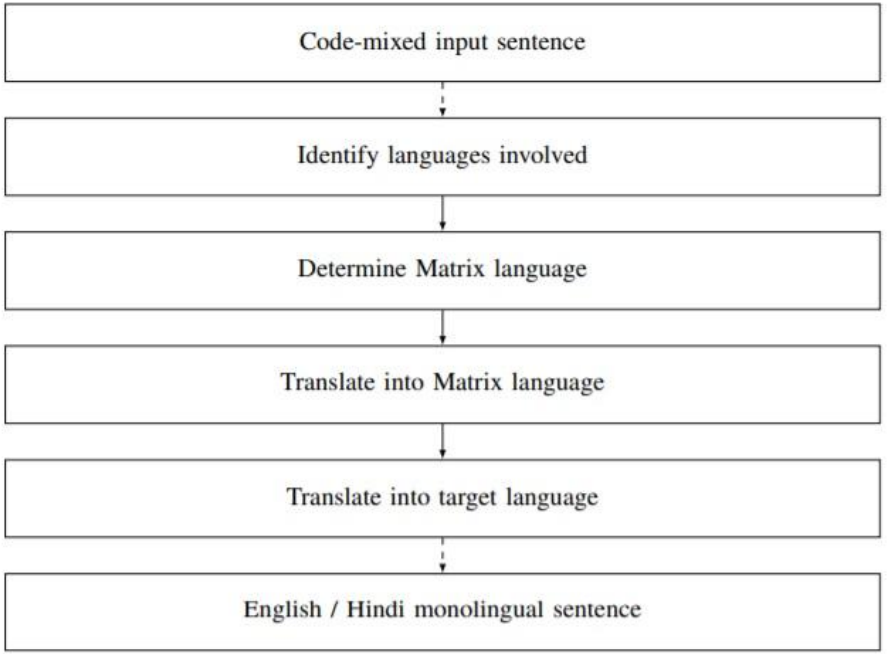
Motivating Example

- [1]The number of english speaking users equals about ¼ of the total users.
- Has gained popularity on Social media and communication platforms like Whatsapp, Facebook, Twitter etc.
- In India, certain English words have almost completely replaced their Hindi counterparts in the lexicon of its people. For example, ‘Kalam’ - ‘Pen’

1. Source : Internetworldstats.com

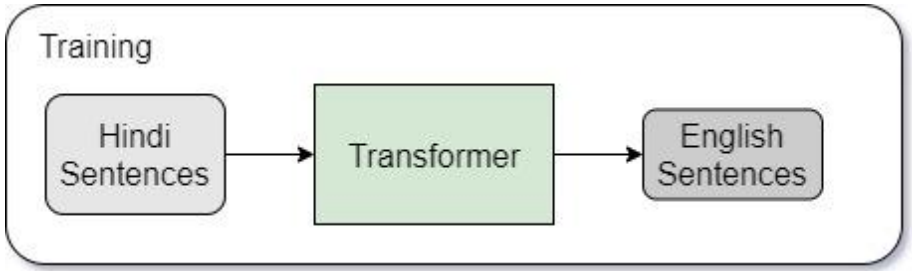
Baseline System

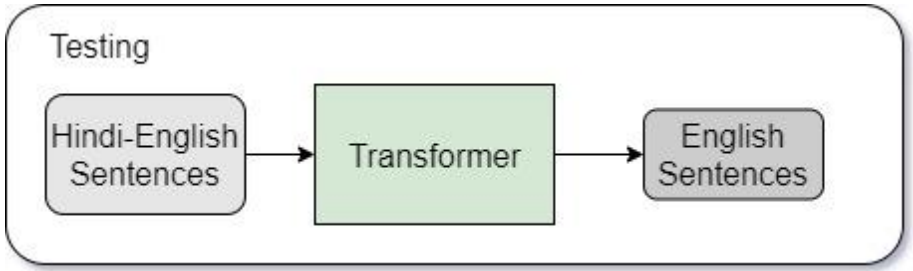
- Existing NMTS + Augmentation
- Augmentation as shown in figure
- Code-Mixed = Matrix + Embedded Language



Improvements - I

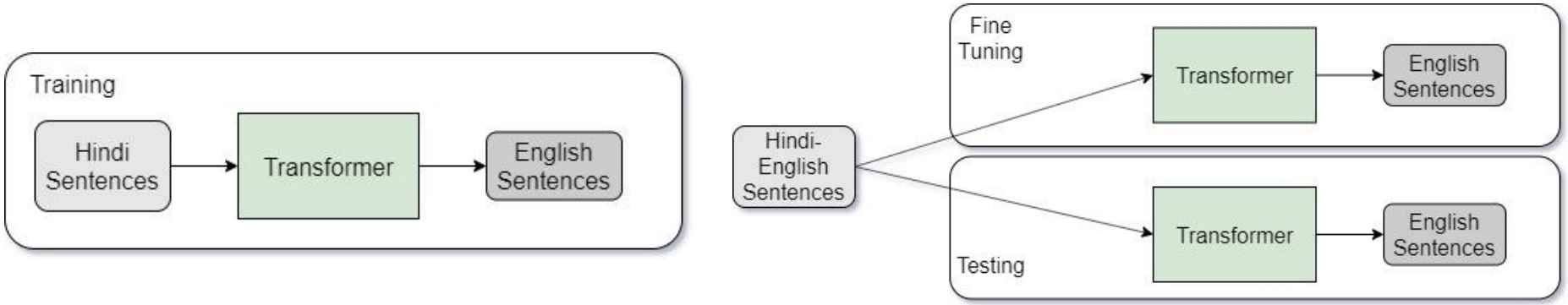
- Baseline utilized Google NMTS (seq2seq) to translate augmented code-switched sentences.
- Custom transformer model was trained using Monolingual Hindi-English parallel corpora.
- Code-Switched sentences were translated using the custom transformer model.





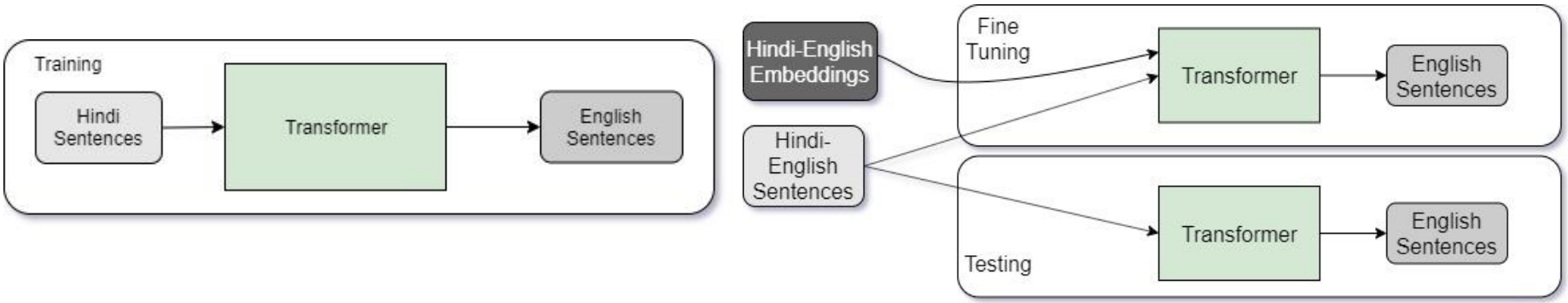
Improvements - II

- Previously trained Transformer model was fine tuned using Hinglish-English parallel corpora.
- Fine tuned transformer model was used to translate code-switched sentences



Improvements - III

- Hindi and English word embeddings were trained using Fasttext.
- Monolingual Transformer model was augmented with trained word embeddings and was fine tuned using Hinglish-English parallel corpora
- Fine tuned transformer model was used to translate code-switched sentences



Work in Progress

- Using Convolutional Sequence to Sequence architecture for machine translation
- Dataset augmentation using Hindi to English word map
- Infusing BERT contextualized embeddings into the translation model
- Mixed model tuning using a mix of monolingual and code-switched translation pairs