Report

- Pranav Goyal (2018114014)

Note: The results for this report have been generated in the notebook: 3_ResultAnalysis.ipynb

Test Sentences:

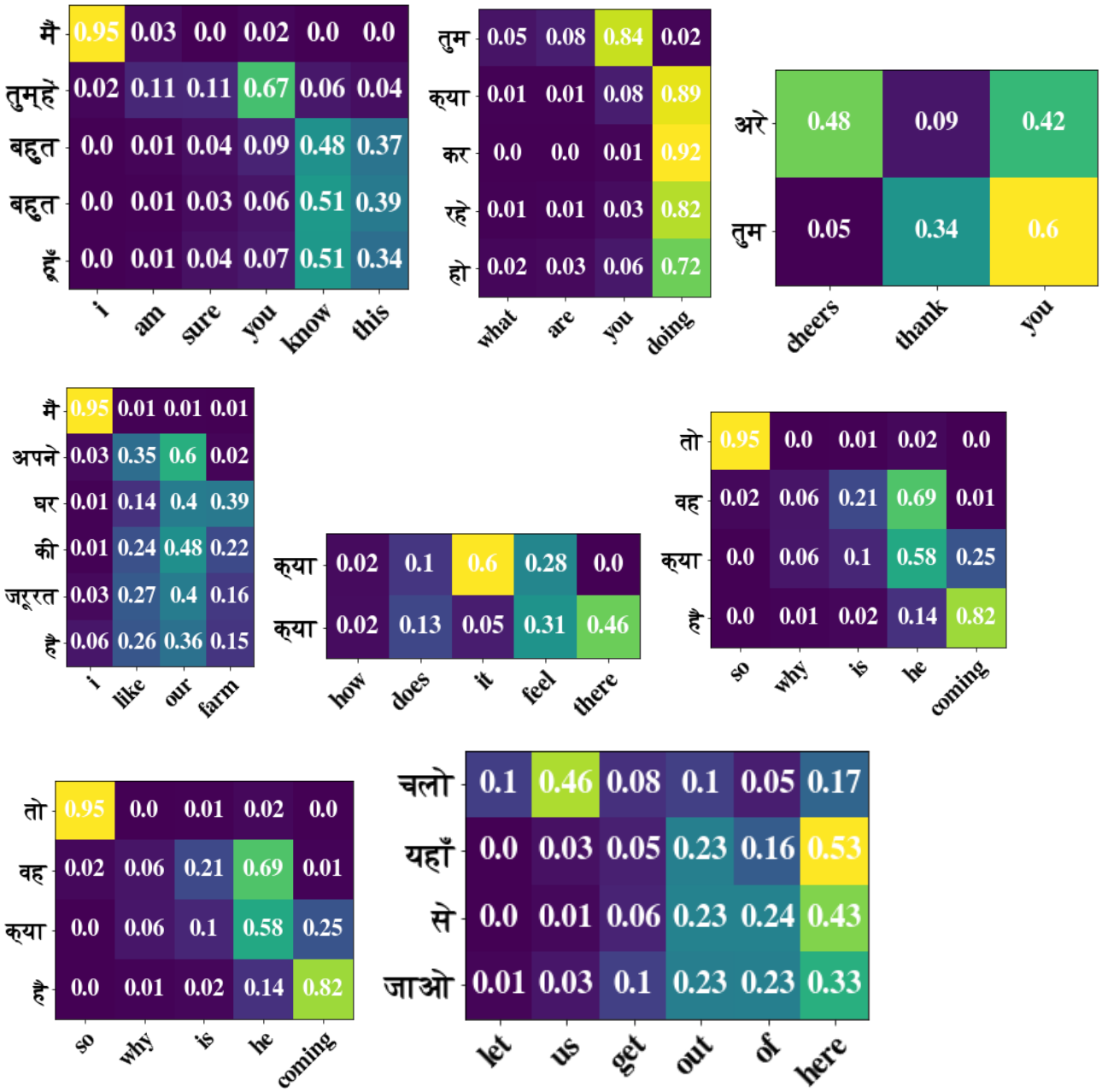
cleaned_english	cleaned_hindi	generated_translation	generated_translation_attention
[i, am, sure, you, know, this]	[<begin>, मैं, यकीन, है, कि, आप, यह, पता, कर, ...	[<begin>, मैं, तुम्हें, पता, है, कि, यह, है, <...	[<begin>, मैं, तुम्हें, बहुत, बहुत, हूँ, <end>]
[what, are, you, doing]	[<begin>, क्या, कर, रहे, हो, <end>]	[<begin>, तुम, क्या, कर, रहे, हो, <end>]	[<begin>, तुम, क्या, कर, रहे, हो, <end>]
[cheers, thank, you]	[<begin>, सज़नों, धन्यवाद, <end>]	[<begin>, ओह, तुम, <end>]	[<begin>, अरे, तुम, <end>]
[i, like, our, farm]	[<begin>, मुझे, हमारे, खेत, पसंद, हैं, <end>]	[<begin>, मैं, बहुत, की, तरह, <end>]	[<begin>, मैं, अपने, घर, की, जरूरत, है, <end>]
[how, does, it, feel, there]	[<begin>, यह, कैसे, वहाँ, लग, रहा, है, <end>]	[<begin>, यह, यह, क्या, है, <end>]	[<begin>, क्या, क्या, <end>]
[so, why, is, he, coming]	[<begin>, तो, क्यों, आ, रहा, है, <end>]	[<begin>, वह, है, कि, क्या, है, <end>]	[<begin>, तो, वह, क्या, है, <end>]
[you, are, going, to, be, okay]	[<begin>, तुम, ठीक, हो, जा, रहे, हैं, <end>]	[<begin>, तुम, यहाँ, जा, रहे, हो, <end>]	[<begin>, तुम, यहाँ, जा, रहे, हो, <end>]
[let, us, get, out, of, here]	[<begin>, चलो, यहाँ, से, चले, जाओ, <end>]	[<begin>, यहाँ, यहाँ, से, जाओ, <end>]	[<begin>, चलो, यहाँ, से, जाओ, <end>]
[that, bitch, is, gigantic]	[<begin>, उस, कुतिया, से, बड़ा, है, <end>]	[<begin>, यह, क्या, है, <end>]	[<begin>, कि, क्या, है, <end>]
[what, do, we, do, templeton]	[<begin>, हम, क्या, करें, टेम्पलटन, <end>]	[<begin>, क्या, क्या, किया, <end>]	[<begin>, क्या, कर, सकते, हैं, <end>]
[[, wind, blowing,]]	[<begin>, [हवा, बह], <end>]	[<begin>, <end>]	[<begin>, कराह, <end>]
[i, am, just, saying]	[<begin>, मैं, सिर्फ, इतना, कह, रहा, हूँ, <end>]	[<begin>, मैं, यहाँ, में, रहा, हूँ, <end>]	[<begin>, मैं, यहाँ, हूँ, <end>]
[little, pussy, big, knife]	[<begin>, छोटा, चूत, बड़ा, चाकू, <end>]	[<begin>, अच्छा, तरह, <end>]	[<begin>, एक, ©, पीआर, एक, nderएम, <end>]
[still, working, on, the, name]	[<begin>, लेकिन, मुद्दा, यह, है, बढ़ती, ज्वार,...	[<begin>, अब, पर, पर, कर, रहे, हैं, <end>]	[<begin>, अब, पर, पर, हो, रहा, है, <end>]
[you, have, got, no, powers]	[<begin>, आप, कोई, शक्तियां, मिल, गया, है, <end>]	[<begin>, आप, कुछ, भी, नहीं, है, <end>]	[<begin>, आप, कुछ, भी, नहीं, है, <end>]
[why, are, not, you, texting, me]	[<begin>, आप, मुझे, मैसेज, क्यों, नहीं, कर, रह...	[<begin>, तुम, तुम, नहीं, कर, रहे, हो, <end>]	[<begin>, तुम, मुझे, नहीं, कर, रहे, हैं, <end>]
[a, warrior, the, world, salutes]	[<begin>, a, warrior, the, world, salutes, <end>]	[<begin>, एक, एक, की, तरह, <end>]	[<begin>, एक, एक, की, एक, <end>]
[what, are, you, doing]	[<begin>, तुम, क्या, कर, रहे, हो, <end>]	[<begin>, तुम, क्या, कर, रहे, हो, <end>]	[<begin>, तुम, क्या, कर, रहे, हो, <end>]
[are, you, writing, this, down]	[<begin>, तुम, यह, लिख, रही, हो, <end>]	[<begin>, आप, इस, पर, पर, कर, रहे, हैं, <end>]	[<begin>, तुम, क्या, कर, रहे, हैं, <end>]
[so, far, that, is, it]	[<begin>, अब, तक, कि, यह, है, <end>]	[<begin>, यह, यह, क्या, है, <end>]	[<begin>, तो, यह, यह, है, <end>]

The columns represent:

- Cleaned_english: Original English Sentence
- Cleaned_hindi: Target Hindi Sentence from the dataset
- Generated_translations: Translation output of first model (seq2seq)
- Generated_translations_attention: Translation output of second model (seq2seq with attention)

The test translations were generated entirely from the english input. The models require the previous inputs, so to generate the translations, I began the sentence with <begin> and used it to predict the next word which then was appended to the input and so on.

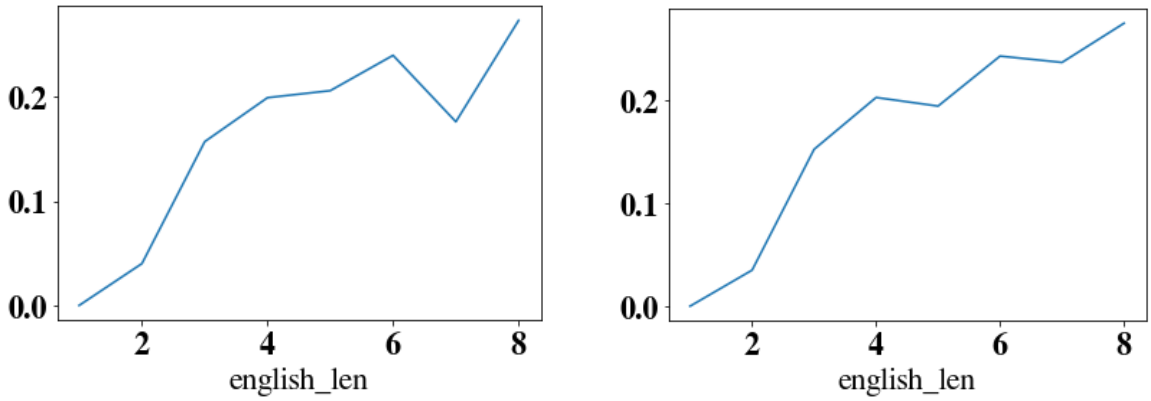
For analysis, I have also displayed the attention matrix from the second model for some of the sentences:



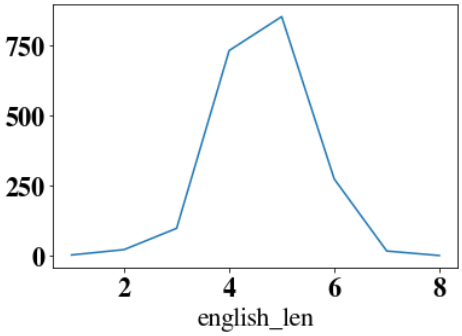
Key Points from the translations and attention matrix:

- Most of the tokens in generated outputs of both models are words of high frequency. None of the uncommon words appear in the generated translations. This is because of the smaller dataset, and use of softmax over the entire list of possible tokens. Uncommon words appear very less in the training dataset and thus do not get assignment much importance, as it is safer for the model to predict a common word instead. The output sequences have limited vocabulary.
    - An improvement for this would be to have pre-trained embeddings for all the token and use the model to predict the embeddings instead, where the closest word to the predicted word would be chosen.
    - An example of this is ‘texting’ missed out in the translation of ‘why are you not texting me’ to ‘tum mujhe nahi kar rhe ho’
- One interesting result for me is that most of the translations are syntactically correct. And they follow the usual word order of SOV. This is actually the result of rnns in both the model, which use the previous output to predict the current output. Even though no embeddings were provided for the hindi input sequence, the models have a sense of what word belongs to what class.
- Attention Matrix show areas of high-confidence are actually words that are very related. ‘So’ and ‘to’, ‘here’ and ‘yahan’, ‘I’ and ‘mai’, ‘you’ and ‘tum’.
- These attention matrices can even be used to generate a phrase translation table. For example in sentence 9, “out of here” and “yahan se jao” which have distributed attentions within the region.

- Some form of overfitting is also visible in the attention matrices, as in the second example, “kya kar rhe ho” has all the attention on the word “doing”.
- Comparing the two models, both show similar results in the test sentences. Among the 2000 entries of test sentences, seq2seq model had a mean BLEU score of 0.203 while the one with attention had a score of 0.200. The attention model is expected to capture long-distance relationships much better as every timestamp of output looks at full encoded sequence and thus has a constant distance to any input. This however is not the visible the test dataset, primarily because all the sentences were of small size (<=8) and both the models performed equally. I also had used smaller hidden state size in attention model (100 as compared to 300 in seq2seq model) as it had much more trainable parameters for the same hidden state size due to the attention mechanism.
- The models have performed really well in some of the sentences by capturing the sense of the input. An interesting case is “I like our farm” to “mai apne ghar ki jarurat hai”. Here, although not close to actual output, the generated results are very close in meaning. You can compare the individual sense of the sentence, ‘i,our’,’mai, apne’ & ‘like’, ’jarurat’ & ’farm’, ’ghar’.
- We can se ‘ek ek’ in the results of both models, which could be because of its abundance in the training set due to it being a common reduplicative.
- The dataset had very little variation in terms of length of the sequences. The following shows the average bleu score for different input lengths. Seq2seq on left, attention model on right.



  - This, at first sight, might seem like an unexpected result as we would expect the bleu score (a measure of accuracy) to be lesser of longer sequences. But actually this is a result of less variation in the original limited dataset. The plot below shows the number of entries of different length. We can see most of the sentences are in the range of [4,6] which isn’t big enough for us to draw any conclusions.



- Both the models were trained for 10 epochs for a batch size of 20 sentences. The loss curve suggests that they both were not a victim of overfitting. (seq2seq on left, attention on right)