Machine Translation

Anushree Korturti
2018102028

Translation: English -> Hindi
Dataset : 100k parallel sentences (70k for train set; 30k for test set)
Epochs for training: 20
Framework: Pytorch
Trained Embeddings with model
Used teacher enforcing for training

Batches size: 32
Loss function: Negative Log Likelihood function from the pytorch library is used.
Optimizer: Adam was used for training of the model.(ignored for the
starting tag and penalized for a longer sentence)

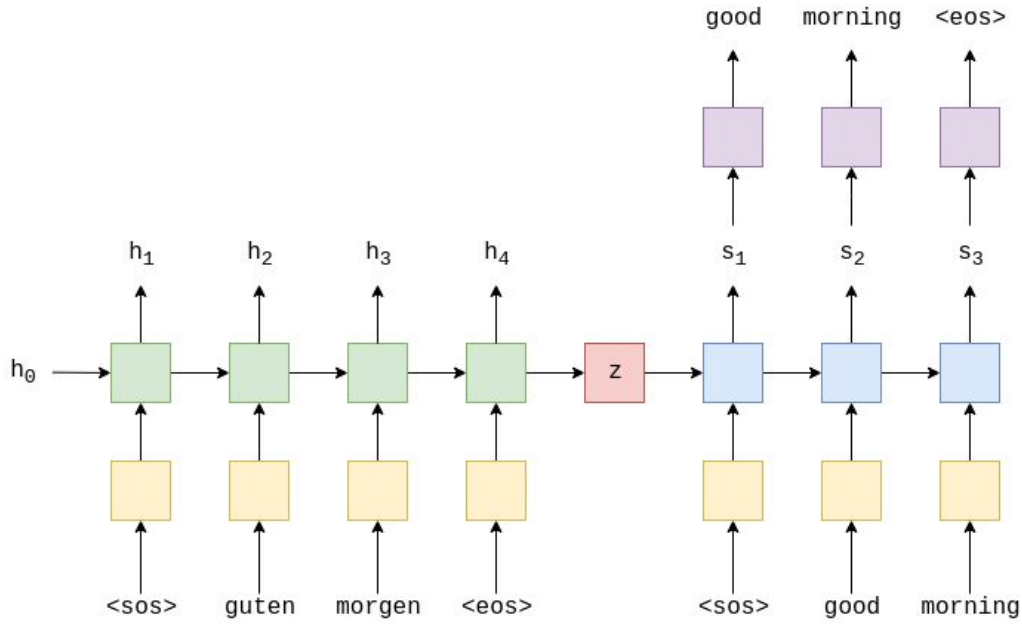
Preprocessing:

1. Expanded all common short forms in eng (e.g. i'm -> i am)
2. Removed any extra spaces.

   Implemented two variants:
   1. With punctuation - Removed all extra characters(including nums, apart from punctuation) in both the English and Hindi corpus.
   2. Without punctuation - Removed all extra characters in both the English and Hindi corpus.

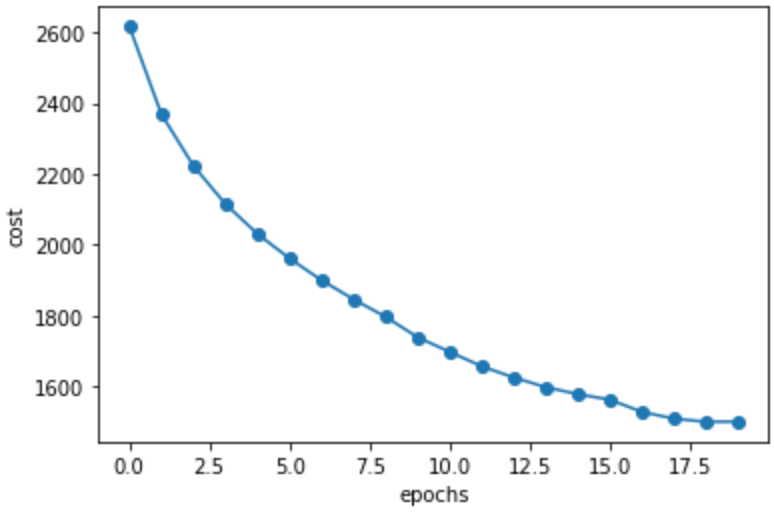
Models:

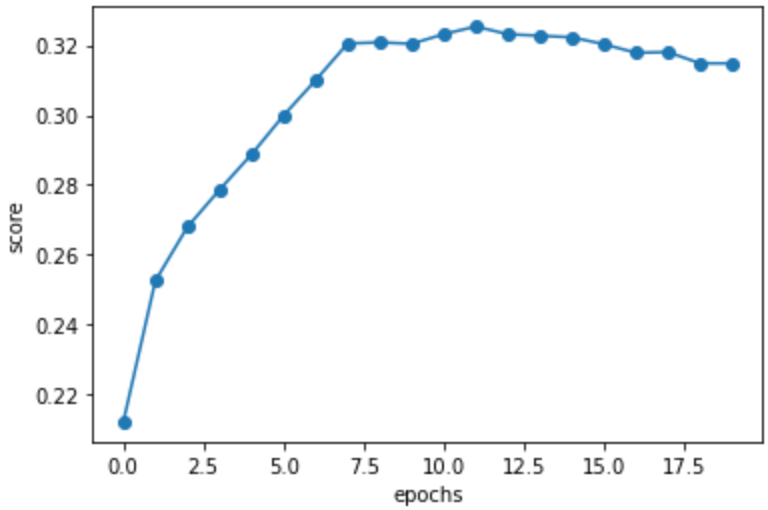
1. Implemented Seq2Seq architecture in the paper ‘Sequence to Sequence Learning with Neural Networks, Stuskever et al’.



Yellow layer: Embedding layer (separate embeddings for target and source language)
Green Layer: Encoder (LSTM)
Z : context vector
Blue layer: Decoder (LSTM)
Purple Layer: Word Predictor

Bleu score = 0.32537157479169493





1. Input: babe , keep walking .
   Target: बेबे , चलते रहो ।
   Output: ठीक है , अपने जाओ ।

2. Input: so , lets go home .
   Target: तो , चलो घर चलते हैं।
   Output: चलो , चलो , चलो ।

3. Input: you can turn here .
   Target: आप यहाँ बदल सकते हैं ।
   Output: तुम यहाँ यहाँ जा रहे हैं।

4. Input: these are my terms .
   Target: ये मेरे शब्द हैं ।
   Output: ये मेरे पास जहाँ हैं ।

5. Input: no , of course not .
   Target: नहीं , बिल्कुल नहीं ।
   Output: नहीं , कोई भी नहीं है।

6. Input: look at those weve conquered .
   Target: हम पर विजय प्राप्त की है उन को देखो ।
   Output: वे हमें बारे कर सकते हैं ।

7. Input: youll never get big .
   Target: तुम कभी बड़े नहीं होगे ।
   Output: अब तुम नीचे सकते हैं ।

8. Input: grandma , how old are you ?
   Target: दादी , कितने साल की हो तुम ?
   Output: अरे , वह क्या कर रहे हैं ?

9. Input: you got it ?
   Target: तुम समझ गए ?
   Output: आप यह करना चाहिए ?

10. Input: what have you done ?
    Target: तुम क्या किया है ?
    Output: क्या किया ?

11. Input: we are on final approach .
    Target: हम क़रीब पहुंच रहे हैं।
    Output: हम पर पर करने के लिए कुछ हैं ।

12. Input: with the secretary of treasury ?
    Target: राजकोष के सचिव के साथ ?
    Output: वहाँ से बात के लिए ?

13. Input: the office is closed tomorrow .
    Target: कार्यालय में कल बंद कर दिया है ।
    Output: के लिए , मैं से ले रहा था ।

14. Input: this is real shot !
    Target: ये सच है !
    Output: यह सब एक मिनट है !

15. Input: do not forget your clips .
    Target: अपने क्लिप को मत भूलना ।
    Output: अपनी को नहीं की तरह ।

16. Input: do i eat others brains ?
    Target: मैं किसी का दिमाग खाता हूँ ?
    Output: मैं कुछ भी हूँ ?

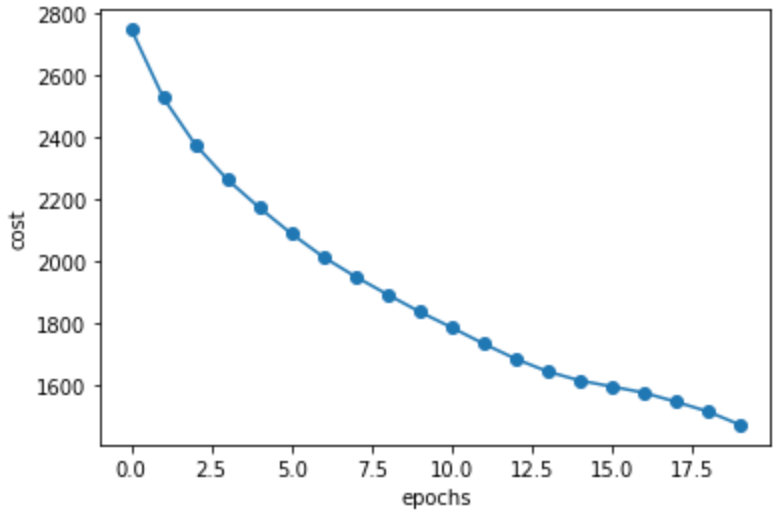
17. Input: they will come for me .
    Target: वे मेरे लिए आ जाएगा।
    Output: वे मेरे लिए देखो ।

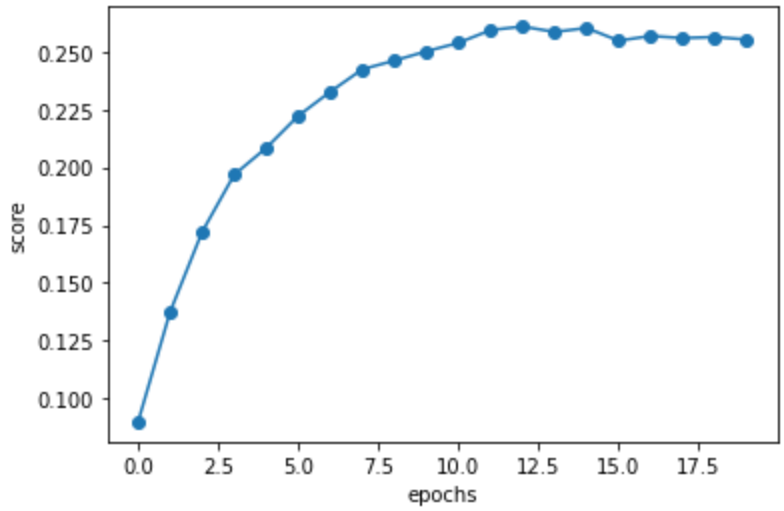
18. Input: you are a goner !
    Target: आप एक प्रबंधक कर रहे हैं !
    Output: तुम एक हो !

19. Input: its not a good time .
    Target: अभी ठीक समय नहीं है।
    Output: यह एक अच्छा नहीं है ।

20. Input: its in the back .
    Target: यह पीठ में है ।
    Output: यह पीठ में है ।

Without Punctuation:
Bleu score (removed punctuation) = 0.26079331461337285





1. Input: are you fine dear
   Target: आप ठीक है प्रिय हैं
   Output: आप ठीक है हैं

2. Input: what brings you down here
   Target: तुम्हें पता है क्यों
   Output: तुम यहाँ क्या हो

3. Input: i can not hear you
   Target: मैं तुम्हें सुन नहीं सकता
   Output: मैं तुम्हें नहीं सकता

4. Input: im not feeling too well
   Target: मैं बहुत अच्छा महसूस नहीं कर रहा हूँ
   Output: मैं एक नहीं बात हो

5. Input: i love you uliya
   Target: मुझे तुमसे प्यार है यूलिया
   Output: मैं तुमसे प्यार हूं

6. Input: are you kidding me
   Target: क्या आप मेरे साथ मजाक कर रहे हैं
   Output: क्या तुम मुझे हो रहे हो

7. Input: you fight for them today
   Target: आज आप उनके लिए लड़ते हैं
   Output: तुम उन्हें के लिए आप के लिए

8. Input: i ca not leave him
   Target: मैं उसे नहीं छोड़ सकते
   Output: मैं उसे नहीं चाहता

9. Input: i will arrest you now
   Target: मैं तुम्हें अब गिरफ्तार करेगी
   Output: मैं तुम्हें आप के लिए होगा

10. Input: i just want to
    Target: मैं बस चाहता हूँ
    Output: मैं सिर्फ में होगा

11. Input: wait wait hold on
    Target: रुको रुको रुको
    Output: रुको रुको रुको

12. Input: get out of here
    Target: यहाँ से चले जाओ
    Output: यहाँ से यहाँ जाओ

13. Input: how did you survive
    Target: आप कैसे बचीं
    Output: तुम कैसे हो

14. Input: after seven years people wonder
    Target: सात साल बाद लोगों को आश्चर्य
    Output: दो पहले के बारे में कुछ जाओ

15. Input: no no i feel good
    Target: नहीं नहीं मैं अच्छा लग रहा है
    Output: नहीं मैं नहीं आ रहा हूँ

16. Input: look at all of us
    Target: हम सभी को देखो
    Output: हमारे पर सब का ले जाओ

17. Input: no of course not
    Target: नहीं बिल्कुल नहीं
    Output: नहीं नहीं

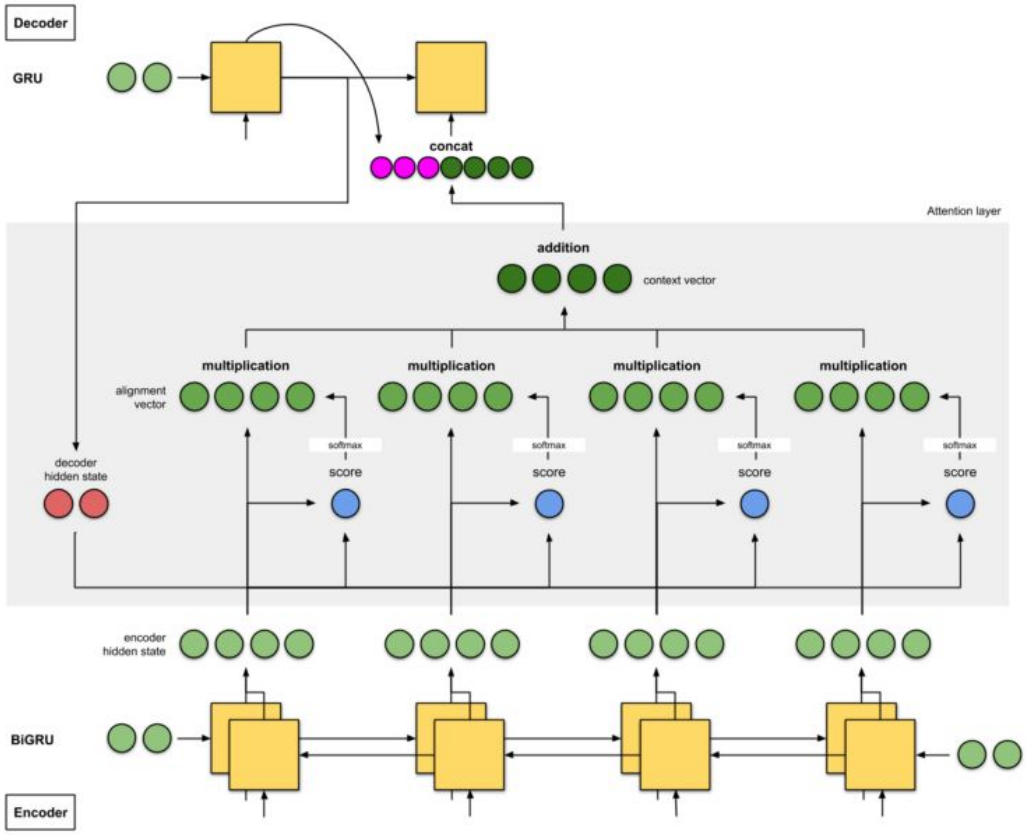
18. Input: thats a good thing
    Target: यह एक अच्छी बात है
    Output: यह एक अच्छा अच्छा है

19. Input: see im good
    Target: देखो मैं अच्छा हूँ
    Output: मुझे अच्छा हूँ

20. Input: we have to help him
    Target: हमें उसकी मदद करनी होगी
    Output: हम हमें मदद कर सकते हैं

Upon removing the punctuation, the translations are better. However, bleu scores go down.

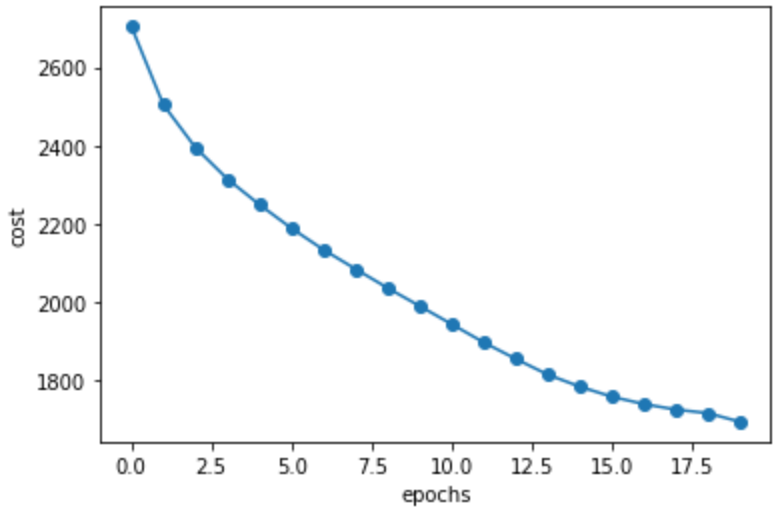
2. Implemented the concept of attention as introduced in Neural Machine translation in the paper ‘Neural Machine Translation by Jointly Learning to Align and Translate,Bhadnau et al.’

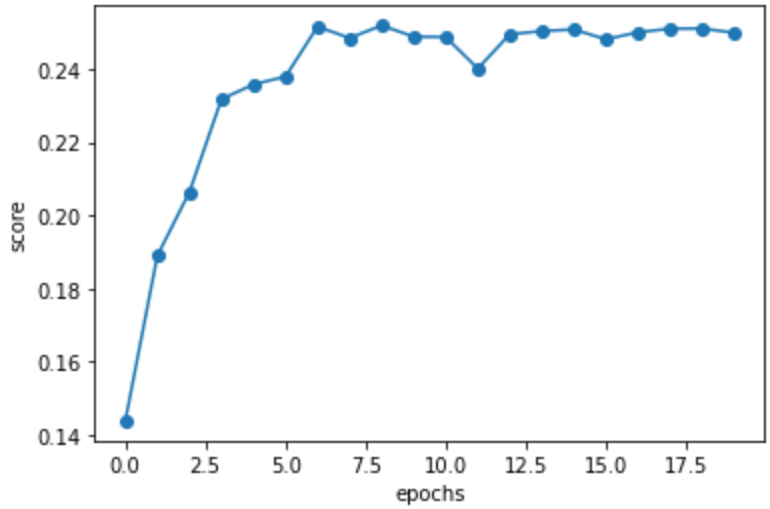


Added attention to earlier model by adding an attention layer and modified decoder to match the architecture.

$$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$$

Bleu score: 0.25189805513789515





1. Input: man speaking native language
   Target: पुरुष अपनी मातृभाषा में बोल रहा है
   Output: पुरुष अपनी में में बोल है

2. Input: what a brilliant machine .
   Target: क्या एक शानदार मशीन है ।
   Output: बहुत एक की है

3. Input: im gonna return it .
   Target: मैं कर रहा हूँ।
   Output: मुझे नहीं मैं हूँ हूँ।

4. Input: this photo was taken then
   Target: यह फोटो तब लिया गया था

Output: यह वह किया है

5. Input: we do not need this guy .
   Target: हम इस आदमी की जरूरत नहीं है।
   Output: हमें हमें करना चाहिए

6. Input: you just keep a lookout .
   Target: तुम सिर्फ एक तलाश रखना । गंदगी !
   Output: आप तुम एक की करते ।

7. Input: hes very dangerous man .
   Target: वो बहुत ही खतरनाक आदमी है ।
   Output: वह बहुत बहुत है । ।

8. Input: she calls herself adele .
   Target: उसका नाम एडेल है ।
   Output: वह दो है बेटे के है । । । । । । । । । । । । । ।

9. Input: are you new to me ?
   Target: मैं तुमसे पहली बार मिल रहा हूँ क्या ?
   Output: आप मुझे करने लिए रहे ?

10. Input: i feel strong , you know .
    Target: मैं सशक्त महसूस करता हूं , यह तुम्हें पता है
    Output: मुझे है मैं हो है यह है ठीक । ।

11. Input: get out of there .
    Target: बाहर आओ।
    Output: वहाँ से से जाओ

12. Input: my name is brown .
    Target: मेरा नाम भूरा है ।
    Output: मेरा नाम है । ।

13. Input: we should not have broken up .
    Target: हम टूट नहीं करना चाहिए था।
    Output: हम कुछ नहीं । ।

14. Input: lets go to the hospital .
    Target: के अस्पताल में चलते हैं।
    Output: चलो में जाओ । ।

15. Input: you are joking , of course .
    Target: आप निश्चित रूप से मजाक कर रहे हैं ।

Output: तुम हो , बाहर हैं रहे

16. Input: he ca not do that !
    Target: वह ऐसा नहीं कर सकते !
    Output: कि नहीं नहीं ऐसा है

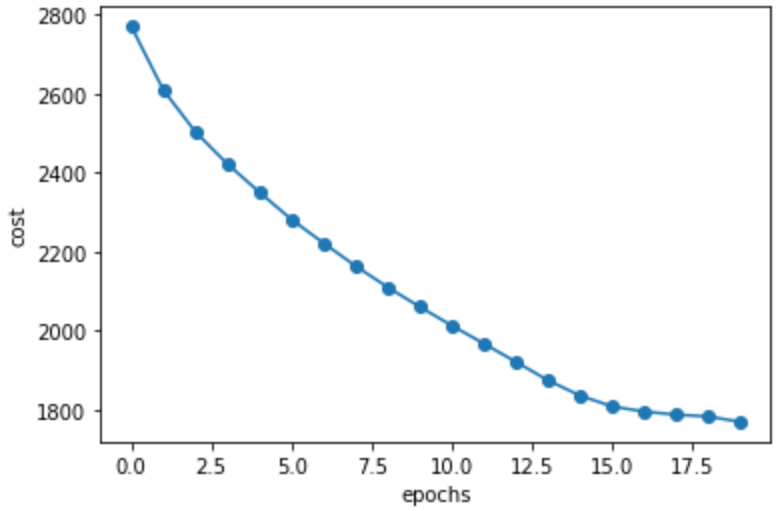
17. Input: is he dead yet ?
    Target: वह मर चुका अभी तक है ?
    Output: वह वह लिए गया है

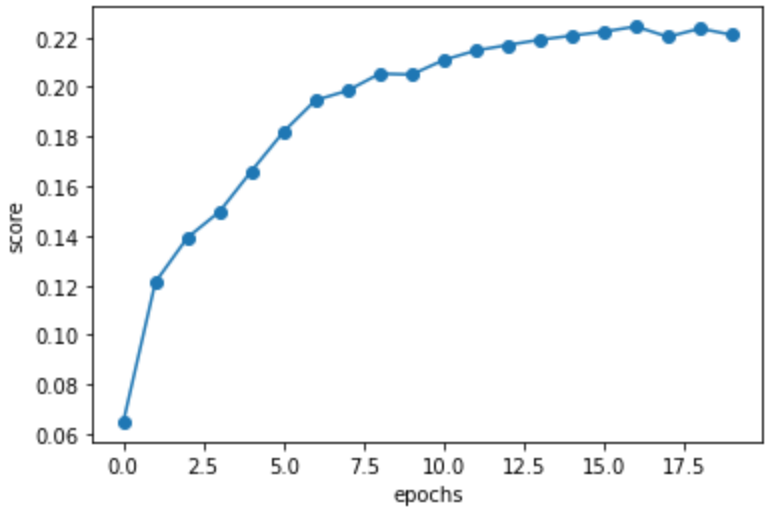
18. Input: the hell it is .
    Target: यह क्या है ?
    Output: इसे है यह है । ।

19. Input: you be careful with that .
    Target: इसे ध्यान से रखना।
    Output: आप भगवान को को किया था

20. Input: i am calling my brother .
    Target: मैं अपने भाई को बुला रही हूँ।
    Output: मैं मेरे हूँ बाहर रहा । ।

Without Punctuation:
Bleu score (removed punctuation): 0.22437762680204032





1. Input: what brings you down here
   Target: तुम्हें पता है क्यों
   Output: यहाँ यहाँ तुम यहाँ

2. Input: wait wait hold on
   Target: रुको रुको रुको
   Output: रुको रुको रुको

3. Input: now i pull a trigger
   Target: अब मैं एक ट्रिगर खींच
   Output: अब मैं उसे

4. Input: it will take you south
   Target: यह तुम्हें दक्षिण ले जाएगा
   Output: आप आप को ले

5. Input: thats a good thing
   Target: यह एक अच्छी बात है
   Output: यह एक अच्छा है

6. Input: so new equals valuable
   Target: तो नए मूल्यवान बराबर होती है
   Output: बहुत बहुत बहुत

7. Input: see im good
   Target: देखो मैं अच्छा हूँ
   Output: मुझे है कि है

8. Input: they ca not stay here anymore
   Target: ये लोग यहाँ अब और नहीं रह सकते

Output: वे यहाँ यहाँ नहीं यहाँ यहाँ

9. Input: thank you baby
   Target: धन्यवाद बेबी
   Output: बहुत धन्यवाद

10. Input: what are you saying
    Target: तुम क्या कह रहे हो
    Output: आप क्या क्या रहे

11. Input: thank you and hurry
    Target: शुक्रिया और जल्दी
    Output: ठीक आप आप को का हैं

12. Input: you call me murderer
    Target: तुम मुझे कातिल कहते हैं
    Output: तुम मुझसे हो

13. Input: if you see anything suspicious
    Target: आप कुछ भी संदिग्ध दिखाई देता है
    Output: आप कुछ भी कुछ भी नहीं चाहिए

14. Input: shell keep you well fed
    Target: वह आप अच्छी तरह से तंग आ रखेंगे
    Output: फिर है तुम्हें तुम्हें पसंद किया

15. Input: let me see it
    Target: मुझे यह देखते हैं
    Output: मुझे मुझे

16. Input: i could be wrong
    Target: मैं गलत भी हो सकता हूं
    Output: मैं कर हूँ

17. Input: makes all the difference
    Target: सभी फर्क नहीं पड़ता
    Output: सभी श्री सभी

18. Input: this is the car
    Target: यह कार है
    Output: यह लिए है

19. Input: ivan had photos too
    Target: इवानभीतस्वीरेंथा

Output: उसने भी याद की लिया है

20. Input: ill figure something out
    Target: मैं कुछ हल निकाल लूँगा
    Output: मैं कुछ मिल गया है

Upon removing the punctuation, the translations make more sense. However, bleu scores go down.
The words seem to repeat more often. This is starkly visible in the with punctuation case, where the punctuation marks repeat.

Although it is expected that with attention the bleu scores are expected to increase, the bleu scores have decreased with attention (both with and without punctuation). This is likely because the sentences are too short for attention to work properly and this is reflected in the scores.

With normal seq to seq, the sentence makes more sense than in attention. However, the translated sentences are more related to the target sentence in the attention model. This shows that with attention, more importance is certain words in the input sentence. This will work better than than the regular RNN network when the sentences are longer.