News Translation For English-Hindi Corpus

Nilabja Bhattacharya
2018201036

Gauravdeep Singh Bindra
2018201027

NLP Applications Project
IIIT Hyderabad

***Abstract*— In this project our task was to perform Machine Translation on a English-Hindi News Corpus. We have solved it using SMT and NMT and we presented the BLEU score on each system**

***Index Terms*— SMT, NMT, General, Concat, Dot, Attention**

I. INTRODUCTION

1) SMT: Statistical machine translation is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule based approaches to machine translation as well as with example based machine translation.

2) NMT: Neural Machine Translation has been receiving considerable attention in recent years, given its superior performance without the demand of heavily hand crafted engineering efforts. NMT often outperforms Statistical Machine Translation (SMT) techniques but it still struggles if the parallel data is insufficient like in the case of Indian languages.

3) Attention: The basic idea: Each time the model predicts an output word, it only uses parts of an input where the most relevant information is concentrated instead of an entire sentence. In other words, it only pays attention to some input words.
Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies without regard to their distance in the input or output sequences.

4) MOSES: Moses is a free software, SMT engine that can be used to train statistical models of text translation from a source language to a target language. Moses then allows new source language text to be decoded using these models to produce automatic translations in the target language. Training requires a parallel corpus of passages in the two languages, typically manually translated sentence pairs.

5) Language Modeling:

$$P(e)$$

Before finding p(f—e) we need to build a machine that assigns a probability P(e) to each English sentence e. This is called a language model.

6) N-grams: For computers, the easiest way to break a string down into components is to consider substrings. An n-word substring is called an n-gram. If n=2,]bigram. If n=3, trigram.

7) Translation Modeling:

$$P(f|e)$$

, the probability of a string f given an English string e. This is called a translation model. P(f — e) will be a module in overall f to e machine translation.
When we see a string f, what we need to

consider for e is that how likely it is to be uttered, and likely to subsequently translate to f? We're looking for the e that maximizes

$$P(e) * P(f|e)$$

.

8) Alignment Probabilities: For a given sentence pair: what is the probability of the words being aligned in particular arrangement. For a given sentence pair, the probabilities of the various possible alignments should add to one.

$$P(a|e, f) = P(a, f|e)/P(f|e)$$

$$P(f|e) = \sum P(a, f|e)$$

9) Expectation Maximization Algorithm
   a) Assign uniform probability values for the alignments.
   b) From this we get the expected counts of alignments.
   c) From these expected counts we get the revised probabilities.
   d) Iterate steps 2 and 3 until convergence.

II. LITERATURE SURVEY

These are the state of the art papers in this area that we referenced.

1. "Neural machine translation by jointly learning to align and translate" (Dzmitry Bahdanau,KyungHyun Cho) In this paper the authors conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoderdecoder architecture, and propose to extend this by allowing a model to automatically soft-search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, they achieve a translation performance comparable to the existing state of the art phrase based system on the task of English-to-French translation. In order to address this issue that the performance of a basic encoderdecoder deteriorates rapidly as the length of an input sentence increases, an extension to the encoderdecoder model which learns to align and translate jointly is suggested. Each time the proposed model generates a word in a translation, it soft-searches for a set of positions in a source sentence where the most relevant information is concentrated. The model then predicts a target word based on the context vectors associated with these source positions and all the previous generated target words.

2. Machine Translation with parfda, Moses, kenlm, nplm, and PRO (Ergun Bicici) In this paper they build parfda (parallel feature weight decay algorithms) Moses SMT models for most language pairs in the news translation task. The authors experiment with a hybrid approach using neural language models integrated into Moses. They obtain the constrained data statistics on the machine translation task, the coverage of the test sets, and the upper bounds on the translation results. Parfda parallelize feature decay algorithms (FDA), a class of instance selection algorithms that decay feature weights, for fast deployment of accurate SMT systems. They train 6-gram LM using kenlm and use mgiza for word alignment.

III. RESEARCH METHODS

- Dataset

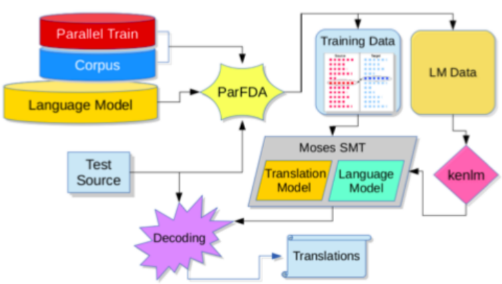
  We have used English-Hindi the parallel training data which consists of the new HindEnCorp, collected by Charles University, linked in the WMT-2014 translation task. The English-Hindi corpus contains parallel corpus for English-Hindi of around 2.7 lakh sentences.

- Data Preprocessing

  We used Moses-toolkit for tokenization and cleaning the English side of the data. The Hindi side of the data is first normalized with Indic NLP library1 followed by tokenization with the same library. As our pre-processing step, we removed all the sentences of length greater than 80 from our training corpus.

- Architecture: SMT

We used a hybrid approach using neural language models integrated into Moses. Obtained the constrained data statistics on the machine translation task, the coverage of the test sets, and the upper bounds on the translation results.
Then trained 3-gram LM using kenlm. In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases, but typically are not linguistic phrases, but phrases found using statistical methods from corpora.
The chosen phrases are further mapped one-to-one based on a phrase translation table, and may be reordered. This table can be learnt based on word-alignment, or directly from a parallel corpus. The second model is trained using the expectation maximization algorithm, similarly to the word-based IBM Model.



$$e_{best} = argmax_e \prod \Phi(\bar{f}_i/\bar{e}_i) d(start_i - end_{i-1} - 1) p_{lm}(e)$$

Score is computed incrementally for each partial hypothesis.
Components :

1) Phrase translation : Picking phrase f to be translated as a phrase e Look up score

$$\phi(f|e\ )$$

from phrase translation table.
2) Reordering : Previous phrase ended in end(i1), current phrase starts at start(i). Compute d(start(i) - end(i-1) - 1)
3) Language model For n-gram model, need to keep track of last n - 1 words

$$Plm(Wi|Wi-\{n\neg1\}, ..., Wi\neg1)$$

for added words Wi

- Training Details: SMT

Corpus Preparation
To prepare the data for training the translation system, we have to perform the following steps:

- Tokenization:
  This means that spaces have to be inserted between (e.g.) words and punctuation.
- Truecasing:
  The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsity.
- Cleaning:
  Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously misaligned sentences are removed.

Language Model Training The language model(LM) is used to ensure fluent output, so it is built with the target language (i.e English in this case).

Training the Translation System
For training we run word-alignment (using GIZA++), phrase extraction and scoring, create lexicalized reordering tables and create your Moses configuration file.

Tuning
Tuning refers to the process of finding the optimal weights for this linear model, where optimal weights are those which maximise translation performance on a small set of parallel sentences (the tuning set). Translation performance is usually measured with Bleu, but the tuning algorithms all support (at least in principle) the use of other performance measures.

- Architecture: NMT

We are using the attention based encoder-decoder architecture. The NMT model consists of an encoder and a decoder, each of which is a Recurrent Neural Network. (RNN)
An encoder neural network reads and encodes a source sentence into a fixed-length vector. A decoder then outputs a translation from the encoded vector. The whole encoderdecoder system, which consists of the encoder and the decoder for a language pair, is jointly trained to maximize the probability of a correct translation given a source sentence.
From a probabilistic perspective, translation is

equivalent to finding a target sentence y that maximizes the conditional probability of y given a source sentence x, i.e.,

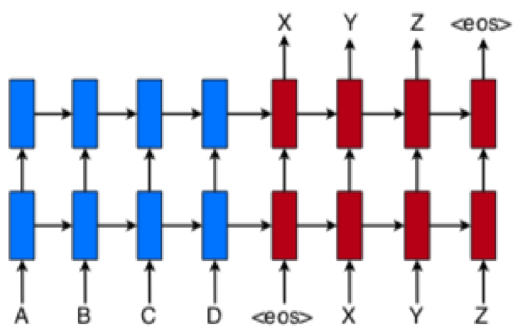
$$argmax_{y} P(y|x)$$

Sequence to Sequence Model

$$h_{t} = sigm(W^{hx} x_{t} + W^{hh} h^{t\text{-}1})$$

$$y_{t} = W^{yh} h_{t}$$

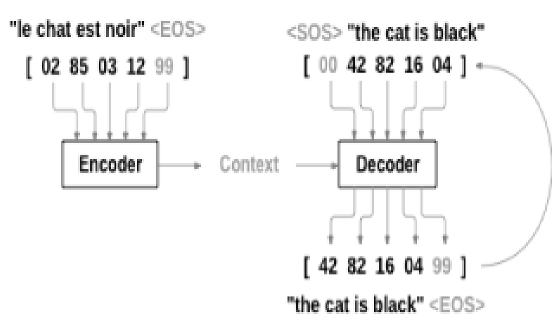
$$P(y_1, ..., y|x_1, ..., x_T) = \Pi P(yt|v, y_1, ..., y_{t-1})$$



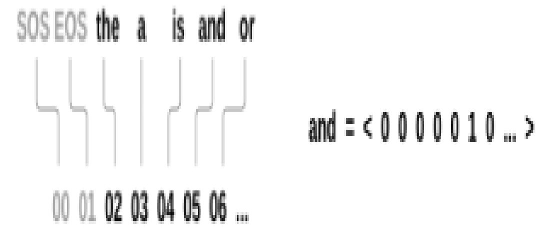
A basic form of NMT consists of two components: (a) an encoder which computes a representation s for each source sentence and (b) a decoder which generates one target word at a time and hence decomposes the conditional probability as:

$$logp(y|x) = \sum_{j=1}^{m} log_p(y_j|y_{<j}, s)$$

- Training details: NMT



Data Processing
We convert the data into one-hot encoding format and then learn the embedding for each word during training.

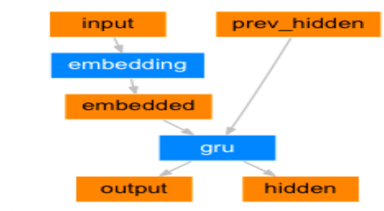


NMT model

In NMT model we have used biLSTM network with layers 2 and 256 hidden unit per biLSTM unit in encoder and decoder. For decoder we have used various type of attention to improve performance

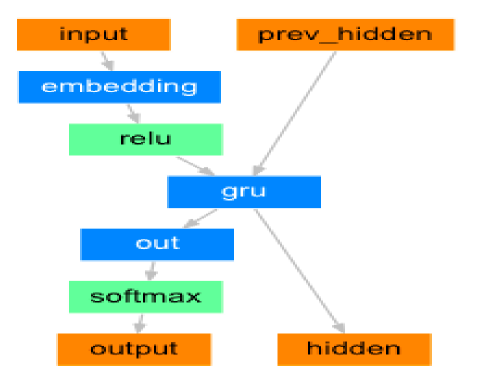
Enoder Unit

We send the hindi sentence through encoder to get a vector space of the sentence. Then this output is passed through the decoder unit to get output.



Decoder Unit

Simple Decoder
In the simplest seq2seq decoder we use only last output of the encoder. This last output is sometimes called the context vector as it encodes context from the entire sequence. This context vector is used as the initial hidden state of the decoder.



Attention based models:
Our various attention-based models are classifed into two broad categories, global and local. These classes differ in terms of whether the attention is placed on all source positions or on only a few source positions.

Attention Decoder
Attention allows the decoder network to focus on a different part of the encoders outputs for every step of the decoders own outputs. First we calculate a set of attention weights. These will be multiplied by the encoder output vectors to create a weighted combination.

Scoring Mechanism

$$\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s] & \textit{concat} \end{cases}$$

Global Attention: The idea of a global attentional model is to consider all the hidden states of the encoder when deriving the context vector ct. In this model type, a variable length alignment vector at, whose size equals the number of time steps on the source side, is derived by comparing the current target hidden state ht with each source hidden state h s:

$$\begin{aligned} \boldsymbol{a}_t(s) &= \operatorname{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \\ &= \frac{\exp\left(\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \end{aligned}$$

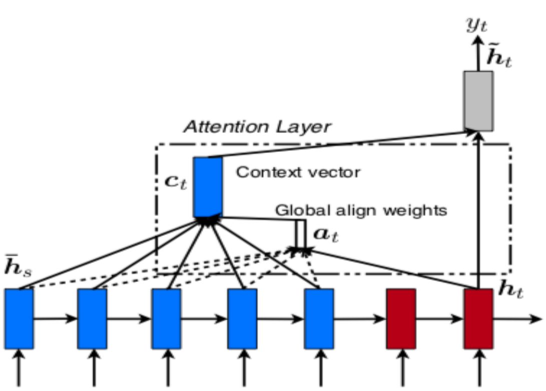


Figure 2: **Global attentional model** – at each time step t, the model infers a *variable-length* alignment weight vector $\boldsymbol{a}_t$ based on the current target state $\boldsymbol{h}_t$ and all source states $\bar{\boldsymbol{h}}_s$. A global context vector $\boldsymbol{c}_t$ is then computed as the weighted average, according to $\boldsymbol{a}_t$, over all the source states.

Local Attention: Our local attention mechanism selectively focuses on a small window of context and is differentiable. This approach has an advantage of avoiding the expensive computation incurred in the soft attention and at the same time, is easier to train than the hard attention approach.

$$\boldsymbol{a}_t(s) = \operatorname{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$

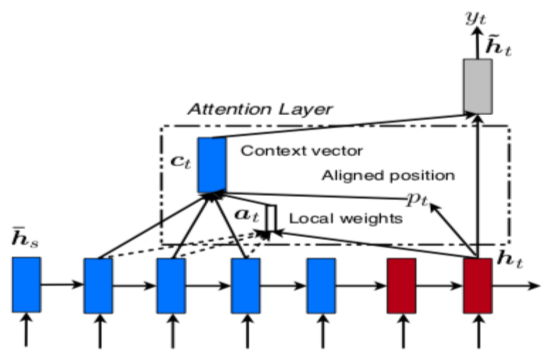


Figure 3: **Local attention model** – the model first predicts a single aligned position p_t for the current target word. A window centered around the source position p_t is then used to compute a context vector $\boldsymbol{c}_t$, a weighted average of the source hidden states in the window. The weights $\boldsymbol{a}_t$ are inferred from the current target state $\boldsymbol{h}_t$ and those source states $\bar{\boldsymbol{h}}_s$ in the window.

IV. FINDINGS AND ANALYSIS

TABLE I
SYSTEM VS BLEU

System(hidden unit, layers, sentence length)	BLEU
SMT	0.2540
Sequence to sequence(256, 2, 100)	0.2577
Concat Attention(256, 2, 100)	0.2492
Dot Attention(256, 2, 100)	0.2574
General Attention(256, 2, 100)	0.2613
General Attention(512, 4, 100)	0.2453
General Attention(512, 4, 25)	0.2824

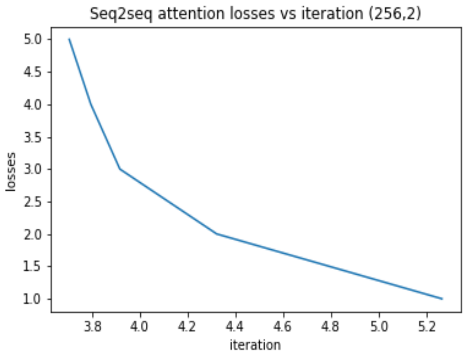


Fig. 1. Plot of loss vs iteration

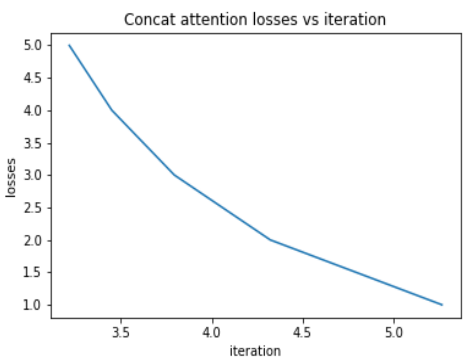


Fig. 2. Plot of loss vs iteration

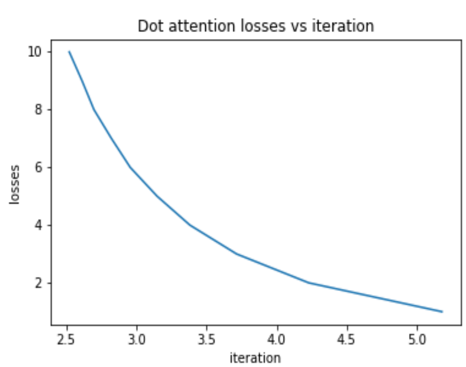


Fig. 3. Plot of loss vs iteration

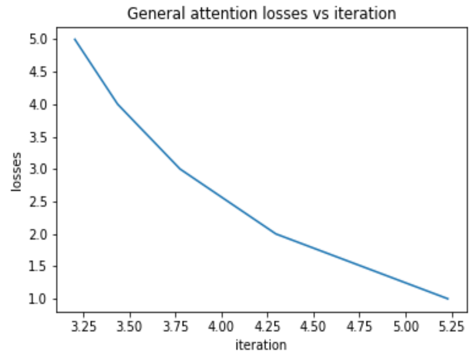


Fig. 4. Plot of loss vs iteration

Analysis and Limitations

- SMT system sentence were not that logical and we had to interpret the translation.
- Many of the hindi words in SMT model were not translated. Some of the prominent cases where this happened was when the word was a common english word written in hindi like stop, rare hindi words like ’changhul’,”benakab” and words which translate to more than one word. Sometimes common words like ”jankari”- information were also not translated.
- There were a lot of punctuation errors in SMT translations.
- Since we have used phrase based trigram model so at max 3 word sentences would be grammaticaly correct
- Increasing LM sentence length would make it too difficult for translating new sentences and it would more algorithmic than machine translation
- We had to deal with sparsity problem in SMT system.
- In sequence to sequence model, the sentence length of generated sentence went upto max length and usually consisted of repeated words.
- There is issue of over and under translation in NMT system.
- General attention based NMT system performed best among all other model.

V. Results

Few results of various models:

A. SMT

```
> Chromoting को बेहतर बनाने में सहायता करना चाहते हैं ?

= want to help improve Chromoting ?

< help improve Chromoting ?


> कुछ किताबों की दुकानें सलाह के पैक बेचतीं हैं जिसमें फार◌ म भी होते हैं .

= some bookshops sell advice packs which includes the forms .

< some bookshops sell advice packs which includes the forms .


> ओसामा बिन लादेन

= Osama bin Laden

< Osama bin Laden


> आयरन मैन २

= iron Man 2

< iron Man 2
```

Fig. 5. Tri-gram phrase based model

B. Sequence to sequence without attention

```
> कटलन डस  कटॉप
= catalan desktop
< catalan desktop machine desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop desktop deskto
p desktop desktop desktop desktop

> लकिन य 400 साल परानी नाल , जिसम पानी बहता ह ,
= but this 400 year old canal , which draws water ,
< but it is this is this is this year this is this year this is this year this is this year this
is this year this is this year this is this year this is this year this is this year this is thi
s year this is this year this is this year this is this year this is this year this is this year
this is this year this is this year this is this year this is this year this is this year this i
s this year this is this year this is this year this is this

> कार  नगी हाल म भी उनका स  वागत समारोह आयोजित हआ जहा उन  होन शिक  षा पर व  याख  यान दिया.उनक प  रससक रथसट ड
निस न उनक विद  यालय क लिए कोष इकट  ठा करन क लए नत  य क कार  यक  रम प  रस  तत किए , हालाकि उनका दश खद ही
आर  थिक-दर  दशा की भयकर चपट म था.कवि न प  राप  त सारी राशि न  ययार  क म बरोजगारो की भलाई क लिए द दी .
= a reception was also organised for him at the carnegie hall where he spoke on education , ruth
st.ddenis who admired him gave dance-recitals to raise funds for his school , but as her own cou
ntry had been recently hard hit by an economic crisis , the poet handed over the proceeds for th
e benefit of the unemployed in new york .
< in lectures regarding the training for a public student in the in in in in in in in in in in i
n in in in in in in in in in in in in in in in in in in in in in in in in in in in in in in in i
n in in in in in in in in in in in in in in in in in in in in in in in in in in in in in in in i
n in in in in in in in in in in in in in in
```

Fig. 6. (hidden unit-256, layers-2, Max sentence length-100)

C. Concat attention based NMT

```
> पिछल हत जब एअर इडिया का विशष विमान कष  णदव राय , जिस पर 16वी सदी क सम  राट का नाम लिखा हा था , प  रधानमत  री
वाजपयी को विएतनाम और इडोनशिया ल गया तो भारत की नई पर  वाभिमख कटनीति म नया मोड आ गया .
= last week , when air-india &apos; s special aircraft , krishna deva raya , flew to vietnam and
indonesia with prime minister atal bihari vajpayee , the name of the 16th century king painted o
n it added an ironic twist to india &apos; s new look-east diplomacy .
< in punjab , the punjab of punjab , punjab , punjab , punjab , punjab , punjab , punjab , punja
b , punjab , punjab , punjab , the punjab of the punjab . the punjab of the punjab . <EOS>

> जिस जररत ह उसक लिय , एन एच एस व  दारा हियरिग एडज , ( श  रवण सहायक उपकरण ) बटरिया और उनकी मरम  मत मफ  त ह
।
= for those who need them , hearing aids batteries or repairs are free on the nhs .
< for example , for example , solvents , , , , , , , , , , , , , <EOS>

> श  री कष  ण न गीता का सन  दश अर  जन को सनाया था ।
= shree krishna told arjun about preachings of gita
< sri krishna krishna krishna was killed by shri krishna <EOS>

> पाठ साफ  कर
= clear text
< clear text <EOS>

> नजीब महफज
= naguib mahfouz
< sonakshi <EOS>

> चीन की लिखित भाषा प  रणाली विश  व की सबस परानी ह जो आज तक उपयोग म लायी जा रही ह और जो कई आविष  कारो का स  रोत
भी ह ।
= the chines written language is the oldest in the universe and is still being used has been use
d in innovative ways
< the language of the is the is the is the language is now is the language and the language <EOS
>

> खोल
= open
< open <EOS>
```

Fig. 7. (hidden unit-256, layers-2, Max sentence length-100)

D. Dot attention based NMT

```
> ( बी ) प  रशिक  षण और कार  य अनभव
= -lrb- b -rrb- training and work experience
< ( applause -rrb- training and training training <EOS>

> और म अरब-इजरायल सघर  ष क सम  बन  ध म हमास की विजय को लकर तटस  थ ह .
= not much separates hamas anti-zionism from fatah anti-zionism except that hamas terrorists spe
ak forthrightly while fatah terrorists obfuscate . even their tactics overlap , as fatah denies
the existence of israel and hamas negotiates with israelis . differing emphases and styles , mor
e than substance , distinguishes their attitudes toward israel .
< and i hamas hamas hamas anti-zionism anti-zionism the hamas the hamas the . . . . <EOS>

> सप  टिसीमिया क  या ह ?
= what is septicaemia ?
< what is the ? ? <EOS>

> स  वर
= vowel
< swara <EOS>

> पासवर  ड रख ( ओई सफ )
= set password ( oi safe )
< encrypt password ( oi safe ) <EOS>

> पर  चो म मसलमानो को नौकरी पर रखन और उनस खरीद-फरोत करन क खिलफ चतावनी दी गई ह .
= they warn against employing muslims , buying from them or selling to them .
< the the often to to and and to to to to to to . . <EOS>

> कब उसकी पलक धीर - धीर हिलन लगी , उस पता नही चला ।
= he did not even notice her eyelashes begin to quiver .
< the had had to to , , to . . . <EOS>
```

Fig. 8. (hidden unit-256, layers-2, Max sentence length-100)

E. General attention based NMT

```
> उस गति स सबधित मीट  रिक जिनस % { short _ product _ name } अनरोधित कार  यवाही निष  पादित करता ह
= metrics relating to the speed with which % { short _ product _ name } performs requested actio
ns
< metrics relating to add % { short _ frame _ frame } from the requested _ frame <EOS>

> उनकी निष   ठा अपन घरो म भ  ट  टियो म आग जलान , अपन दवता क लिए भजन गान और चावल , दध सोमरस या पश बलि क रप
म अर  पित करन म थी .
= their devotion consisted in burning fires in their hearths , singing hymns to their gods and o
ffering rice , milk , soma or animals as sacrifice .
< his speeches in his his , his his his , his , his , his , his , , and and and and and his , <E
OS>

> बाबर और उसक पत  र हमाय दोनो क पास भारत म राजनतिक और सास   कतिक एकता उत   पन   न करन का महान कार  य करन की
व  यापक दष   टि तथा कल   पना थी , कित उनका शासन बहत कम समय तक रहा .
= both babur and his son humayun had the breadth of vision and the imagination to set about the
great task of creating political and cultural unity in india ; but they had very brief reigns .
< the the and the and and and and and and the the the the the the the the the the the , and , ,
, , and and and the . <EOS>

> अगर आपको अपनी स  थिति क बार म कोई शइका हो , तो हस  ताक  षर न कीजिए और घर वापस लौटन तक इतजार कीजिए और उसक
बाद काननी सलाह लीजिए ।
= if you are uncertain about your position you should not sign but wait until you get home and s
eek legal advice .
< if you have the , , your , , , , your , , , , , you and your <EOS>

> उनम स एक म चावल क अतिरिक  त सब कछ हो ; दसरी म दही क अलावा सब कछ हो .
= one of them has all the items except rice , the other has all the items except curds .
< some of them have to to few of , , to of . . . . . . . . <EOS>

> और मझ करीब २०० तरह क दश पता ह , और मझ आकड पता ह ।
= and i know 200 , i know about the small data .
< and i have a small figure of my data , i know . <EOS>
```

Fig. 9. (hidden unit-256, layers-2, Max sentence length-100)

F. General attention based NMT

```
> व टस  ट क  रिकट म सबस ज   यादा रन बनान वाल बल  लबाज  ह ।
= he has scored the maximum runs in test cricket .
< they is the the of the the cricket of . . <EOS>

> pulse aqui para cambiar su apodo
= click to change your nickname
< click to change the nickname <EOS>

> लवासा
= lavasa
< shirdi <EOS>

> changer de mot de passe ...
= change password ...
< change nickname ... <EOS>

> जाचसची
= checklist
< polish <EOS>

> म अपन प  रदाता नही ढढ  सकता ह और म इस दस  ती रप स दाखिल करना चाहता ह ( m ) :
= i can &apos;t find my provider and i wish to enter it manually :
< i can want to want to my my i i i i : <EOS>

> यह सब फ  रिश  त नही ह ,
= all of them are not angels ,
< the all is not not , <EOS>

> सशन कार  यक  रम
= session programs
< marquee survey <EOS>

> आप क  या कर सकत ह अगर आप को शिकायत करनी हो ।
= what to do if you wish to complain
< what if if if if if if you can . <EOS>
```

Fig. 10. (hidden unit-256, layers-2, Max sentence length-100)

G. General attention based NMT

```
> एवोल  यशन बकअप जाच
= check evolution back up
< check evolution address <EOS>

> डिवाइस चयन
= device selection
< the device <EOS>

> नोबल परस  कार जीतन क बाद , वह चार दशको तक गमनामी म जीवन व  यतीत करता रहा ।
= after winning the noble prize , he lived in obscurity for four decades .
< after the father of his noble of the the he the father of his father . <EOS>

> घ  l18 आप को सामाजिक निधी स मदद
= gl18 help from the social fund
< gl18 you tenant you benefit <EOS>

> इडफिनिट
= indefinite
< rwanda <EOS>

> अल  लाह उस गमराह कर दगा ।
= allah will destroy them
< allah will make it misleading . <EOS>

> उन  होन पहल ही हम चतावनी दी थी । &lt; s &gt; यह शरआत ह ।
= they told us what to expect . now it ' s beginning .
< they asked that we lesson . <EOS>

> तो हम एक दसरी धारणा की और कदम बढ ात ह ,
= then we move on to a second assumption ,
< so we have a a a a a , , <EOS>
```

Fig. 11. (hidden unit-512, layers-4, Max sentence length-25)

VI. FUTURE SCOPE

We can work on the following points to improve the performance in the future

- Using Tranformer architecture.
- Having a larger dataset will surely improve performance.
- Applying coverage can prove to be useful.
- Using pre-trained embedding.

VII. CONCLUSIONS

General attention based NMT system performed better as compared to all other models. It was also seen that if max length was small, the model performed relatively better than one with large max length. The BLEU score can be further improved using a transformer model and using large dataset. We can also use coverage to make a better model.

VIII. LINKS

- Slides
- Github
- Complete Repo

REFERENCES

[1] pytorch
[2] moses
[3] Machine Translation with parfda, Moses, kenlm, nplm, and PRO(Ergun Bicici)
[4] "Neural machine translation by jointly learning to align and translate" (Dzmitry Bahdanau,KyungHyun Cho)
[5] Sequence-to-SequenceFile
[6] NMT by Jointly Learning to Align and TranslateFile
[7] Effective Approaches to Attention-based Neural Machine TranslationFile