Assignment 2 - E1 246: Natural Language Understanding (2019)

Apoorv Saxena
apoorvumang@gmail.com

Corpus	Eng/German	Eng/Hindi
Max sentence length	10/17	10/10
Vocabulary	10k/17k	15k/21k
No. of sentences	176,692	93,016
No. of tokens	2,285,670	631,949

Table 1: Statistics of data used

Abstract

The task is Neural Machine Translation using seq2seq models and Attention. We have implemented translation from German to English and Hindi to English using multiple different models and shown performance results and insights obtained through experimentation.

1 Data

Parallel data for German-English translation was taken from Tatoeba.org (Tat) project. Parallel data for Hindi-English translation was taken from OpenSubtitles (Ope) project. Basic text preprocessing was used to tokenize into words.

Translation was done from German to English and Hindi to English since it it is easier to interpret and analyze. 80-10-10 train, validation test split was done.

The statistics for the data can be seen in Table 1.

2 Models

Three different types of models were used for translation. Cross-entropy loss function was used in all experiments.

2.1 seq2seq (Sutskever et al., 2014)

This was the simplest model of the 3. It consisted of

1. An embedding layer (fully connected layer)
2. LSTM encoder layer
3. LSTM decoder layer
4. Fully connected layer

Size of the embedding layer was 256. Both LSTMs had 256 units. This remained the same for all subsequent models. Figure 1 shows this architecture. 10.5M parameters and 8.5M parameters were trained in the hindi-english and german-english tasks respectively.

2.2 Additive attention(Bahdanau et al., 2015)

Additive attention was applied at the decoder. Rest of the model remains the same. The weights for computing the context vector were

$$e_{j,t} = V_a \cdot \tanh(W_a s_{t-1} + U_a h_j)$$

$$\alpha_{j,t} = \frac{\exp(e_j)}{\sum_{k=1}^{T} \exp(e_k)}$$

where s_{t-1} is the decoder cell state at step $t-1$ and h_j is the encoder output at time step j. $\alpha_{j,t}$ are the weights. All other values in the equation are trainable weights.

Figure 2 shows this architecture. 50M parameters were trained in the German-English task.

2.3 Self attention(Vaswani et al., 2017)

Self attention was applied at both the encoder and decoder. Additive attention was used. At both the encoder and decoder, the new output was calculated as follows.

$h_{t,t'} = tanh(x_t^T W_t + x_{t'}^T W_x + b_t)$
$e_{t,t'} = \sigma(W_a h_{t,t'} + b_a)$
$a_t = softmax(a_t)$
$l_t = \sum_{t'} a_{t,t'} x_{t'}$

x_t is the input and l_t are the output vectors. Rest all are trainable params.

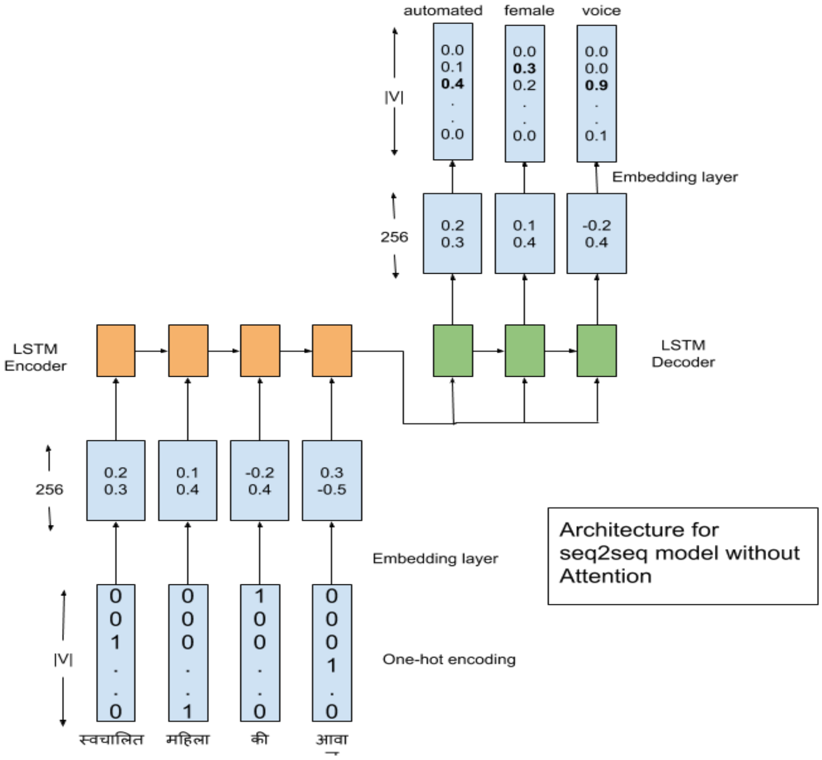

Figure 1: seq2seq model architecture

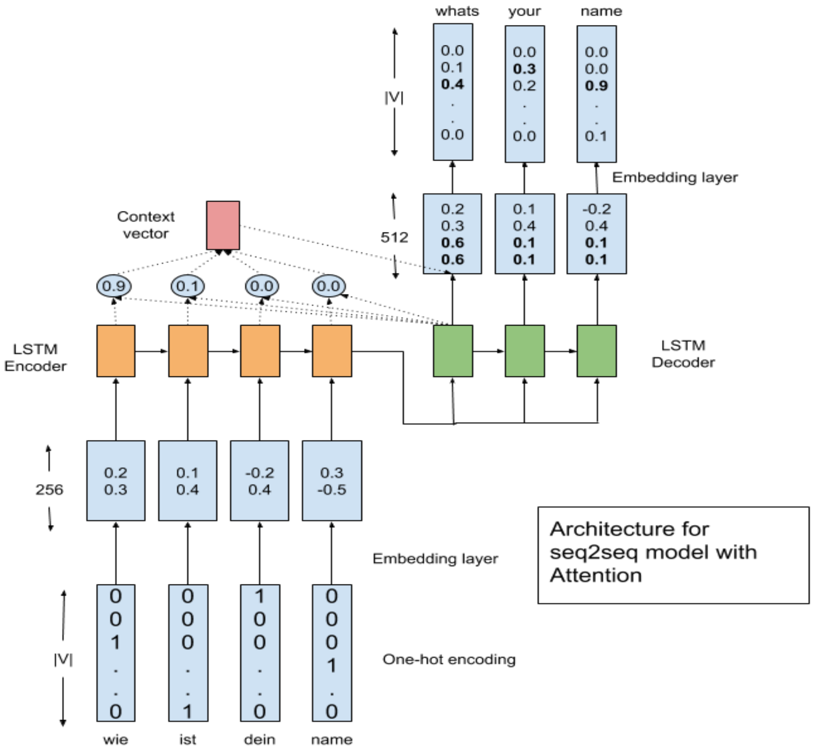

Figure 2: seq2seq model with Additive Attention

	BLEU-1	**BLEU-4**
German seq2seq	0.5812	0.2771
German Attention	0.6404	**0.3501**
German Self attention	0.459	0.078
Hindi seq2seq	0.363	**0.056**
Hindi Self attention	0.276	0.015

Table 2: BLEU scores for different experiments. They have been calculated on the test split of the dataset

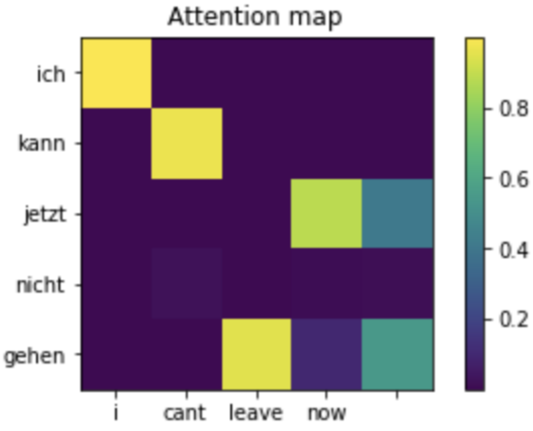

Figure 3: Additive attention

3 Training the models

Each of the above models was trained on a single GPU for 4 hours.

4 Results

In all, we are reporting 5 experimental results on German-English and Hindi-English translation. BLEU scores can be seen in Table 2. Table 3 shows some example translation where our models succeeded and failed. Figures 2-5 show the attention weights corresponding to certain translation (additive attention).

5 Analysis

In figures we can see the attention weights corresponding to different words in the encoder output at each time-step of the decoder. As we can see, our model is able to attend very well to the input words. This can be used as an automated alignment technique.

The best performance was obtained using Additive Attention. BLEU-1 scores correspond to unigram matching, so they are good for all models. BLEU-4 includes n-gram matching upto 4-grams. German-English translation performs much better

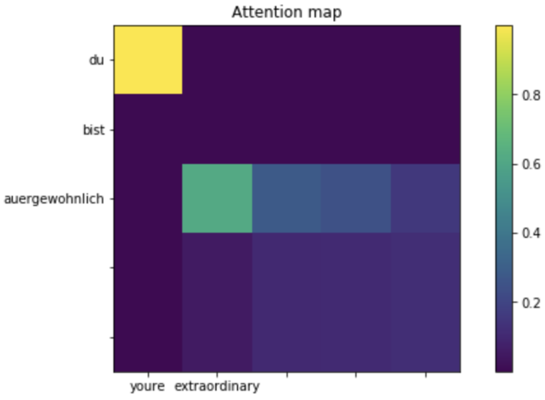

Figure 4: Additive attention

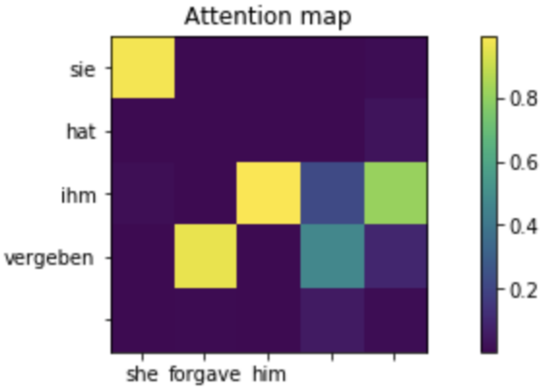

Figure 5: Additive attention

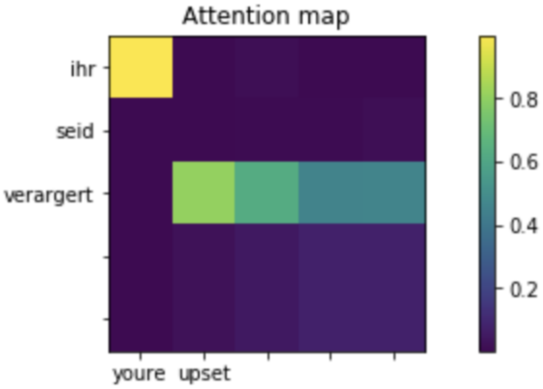

Figure 6: Additive attention

German	Our Translation	Actual Translation
tom hat zwei katzen	tom has two cats	tom has two cats
ich war nicht erfreut	i wasnt pleased	i was not pleased
tom ist klug	tom is clever	toms smart
sie klingen nicht verangstigt	you dont sound	you dont sound scared
das ist ein enormer verlust	its a bad loss	this is a huge loss

Table 3: Additive attention in German-English translation

German	Our Translation	Actual Translation
tom hat zwei katzen	tom has two cats	tom has two cats
ich war nicht erfreut	i wasnt pleased pleased	i was not pleased
tom ist klug	tom is smart	toms smart
sie klingen nicht verangstigt	*you not upset*	you dont sound scared
das ist ein enormer verlust	*thats a a thrill*	this is a huge loss

Table 4: seq2seq German-English translation

German	Our Translation	Actual Translation
maria kann gut tanzen	*mary mary dance dance dance*	mary can dance well
wie ist dein name	what is your name	what is your name
ich hatte einen schrecklichen traum	*i a a a dream*	i had a terrible dream
ich traf tom bei ihm zu hause	*i met tom at him*	i met tom at his house
endlich ist es freitag	*finally finally*	finally its friday

Table 5: Self Attention German-English translation

Hindi	Our Translation	Actual Translation
तुम्हारे पिताजी कहा करते थे:	your father you say	your dad used to say
स्वचालित महिला की आवाज:	automated female voice	automated female voice
मुझे खेद है, वह यहाँ नहीं हूँ.	im not im here here	im sorry hes not here
मैं ठीक तुम्हारे पीछे ही हूँ!	im im right you	im right behind you
यहाँ अपने भोजन है.	heres your your	heres your food

Table 6: seq2seq Hindi-English translation

Hindi	Our Translation	Actual Translation
तुम्हारे पिताजी कहा करते थे:	you you you you	your dad used to say
स्वचालित महिला की आवाज:	alarm alarm	automated female voice
मुझे खेद है, वह यहाँ नहीं हूँ.	im im im	im sorry hes not here
मैं ठीक तुम्हारे पीछे ही हूँ!	im im you	im right behind you
यहाँ अपने भोजन है.	your your your	heres your food

Table 7: Self Attention Hindi-English translation

here compared to Hindi-English translation. This is mainly because of the huge size of vocabulary in hindi compared to training data size. If we increased the corpus size, the vocabulary in hindi increased by a large amount and that made experimenting very hard since it would take over 12hrs to train.

Self attention did not give any noticeable benefits because there weren't much local dependencies in our training data (small sentence lengths) . Additive attention on the other hand decreased training time and increased BLEU scores.

References

OpenSubtitles.com project. http://www.opensubtitles.com. Accessed: 2010-09-30.

Tatoeba.org project. http://www.tatoeba.org. Accessed: 2010-09-30.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.