# Document-Level Machine Translation with Hierarchical Attention

*Experiments with neural machine translation*

Presented to Prof. Chris Pollett
Department of Computer Science
San Jose State University

In Partial Fulfillment
Of the Requirements of the Class
Fall 2022: CS 297

By
Yu-Tang Shen
October 2022

## TABLE OF CONTENTS

# I. Introduction on Neural Machine Translation

As neural networks became dominant in various fields such as visual object detection, trend prediction, etc., researchers tried to utilize neural networks to further improve machine translation. Neural networks demonstrated its ability to resolve complex relationships between inputs and outputs: [1] showed neural networks improved the accuracy of house price prediction from up to seven kinds of input features. It would be difficult for a human being to determine how important each input feature weights, but neural networks can assign the optimal weight to each feature, and perhaps assign weights to mixtures of multiple features.

Although neural networks had presented excelling capabilities, the nature of neural networks might contradict with some tasks. Different from trend prediction and object detection, machine translation has a variable length in both input and output. For example, in a visual object detection application, the input shape is often fixed (28 x 28 in the popular MNIST dataset), and so is the output shape (usually a 10-digit long one-hot vector for MNIST digit classification) [2]. On the other hand, the input shape of a machine translation project can vary from 5 (a short sentence) to 50 (a longer sentence having several clauses in it); moreover, the output shape is also indeterministic: two sentences having same shape can have different length when translated into target language.

Without the capability to consume variable sized input and generate outputs accordingly, it would constrain the machine translation (MT) application to be flexible enough to become practical. A fixed length input and output meant the application could only translate sentences that meet the limitation.

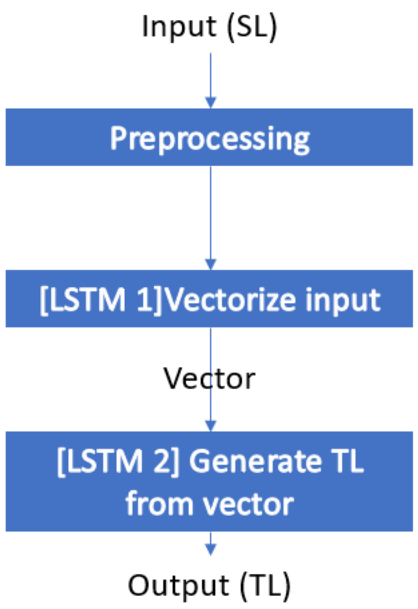
To enable neural networks to consume and output variable shaped sentences, [3] designed a neural network that can generate translations in different shapes as well as consume variable input shapes. Since then, neural machine translation (NMT) has become the dominant technology regarding to machine translation tasks.

## II. Neural Machine Translation Implementation

There are various designs in NMT, and the designs keep updating as this report was written. This report would cover the NMT designs up to when attention technique, and the remaining designs will be covered in the next report.

[3] presented a design that allows neural networks to produce variable shaped outputs. The authors designed a schema that utilized two long short-term memory (LSTM) neural networks. LSTMs are similar to recurrent neural networks (RNNs), but they solved the gradient vanishing problem. Like RNNs, LSTMs take a fixed sized input and generates a fixed sized output.

The authors designed a protocol that two LSTMs could cooperate to give variable length translations as follows:



SL: source language

TL: target language

*Fig 1.*

The inputs were first preprocessed by padding empty strings to the maximum length that this application aims to take so all the inputs had equivalent lengths without introducing significant bias. The first LSTM summarized the input sentences, producing a vector that acted as activations for the next LSTM. The second LSTM then takes the summarized vector and produce the translation until it generates a special token <EOS>, abbreviated from end of sentence.

This method was also referred as encoder-decoder implementation, because the first LSTM essentially encoded the information in the input and the second LSTM decoded the information accordingly.

## III. Experiment

A dual LSTM setup is implemented aiming to translate English sentences into Chinese (shown in the image below). The model was designed with two LSTMs, one for encoding information in English sentence while the other one decodes the information. In order to reduce workload on hardware, an embedding layer was introduced to first reduce the input size: the embedding layer also acted as another encoder to reduce the input size from 15k to 1k. With a reduced size input, the encoder LSTM then summarized the embeddings, and the decoder LSTM eventually generates corresponding output from the summary given by the encoder LSTM. The output at that stage was a reduced vector, which a time distributed layer expands the reduced vector to the original size. Eventually, a softmax layer regularized the values in the outputs and the translation can be recovered from the one-hot outputs. UM-corpus [4] was used to train and evaluate the model.

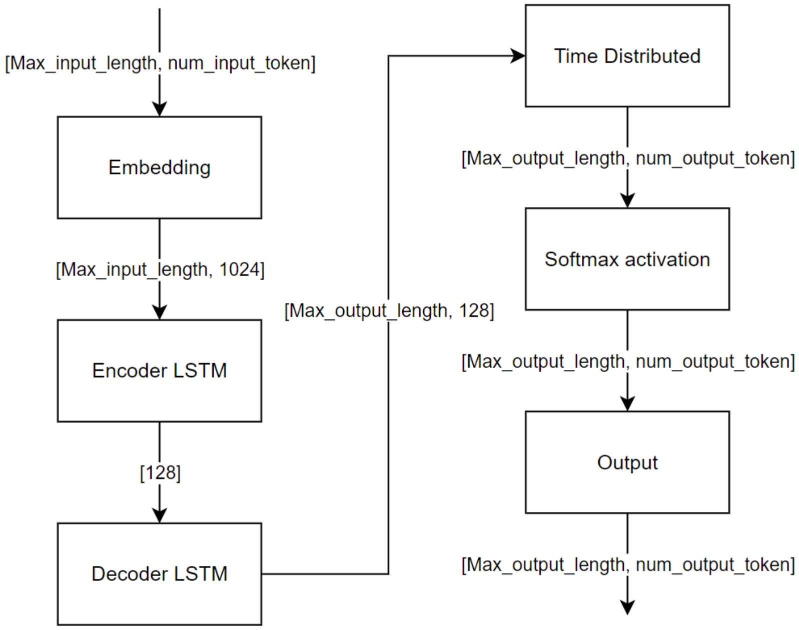

[Max_input_length, num_input_token]
Embedding
[Max_input_length, 1024]
Encoder LSTM
[128]
Decoder LSTM
[Max_output_length, 128]
Time Distributed
[Max_output_length, num_output_token]
Softmax activation
[Max_output_length, num_output_token]
Output
[Max_output_length, num_output_token]

*Fig 2*

With the embedding layer generating a 1024-long vector and configurating both encoder and decoder LSTM to be shape of 128 after 3000 epochs of training, the model gave the accuracy as shown in the image below. The training accuracy kept increasing, but the validation accuracy essentially remained the same. Therefore, the current configuration might not improve further with more training epochs.

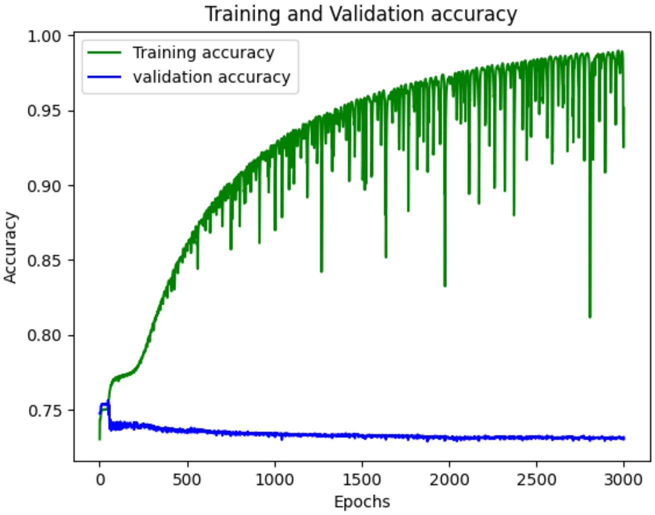
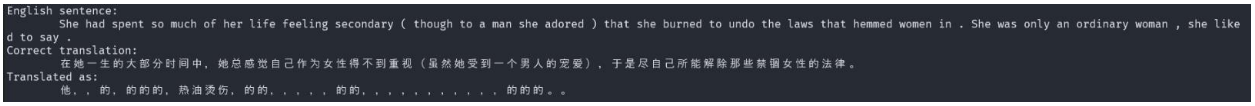


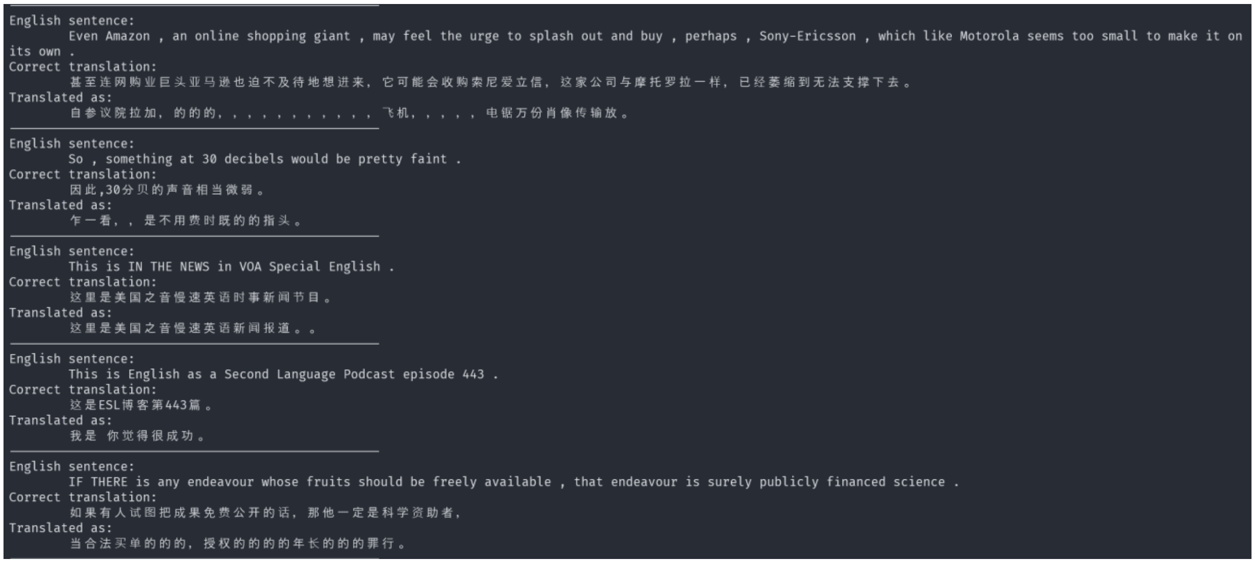
*Fig 3*

Looking at the actual outputs of the model, some sentences were translated properly while some was not. In Fig 5, some results were shown. The first translation completely failed, where the decoder couldn't properly generate any meaningful words but kept repeating the word "的". The reason for getting such result was the encoding for the word "的" was close to the untrained model output (shown in Fig 4), i.e., it could be viewed because of not being trained at all. The second sentence improved on its sentence structure, where less stuttering was observed, but the translation didn't match the meaning of the original sentence. The third translation has a high quality: the model translated the meaning correctly, produced a valid sentence structure, and the difference between the ground truth (时事新闻节目) and the output (新闻报导) was acceptable as they were synonyms. With these three sentences, a gradient of how this model can learn was presented. The model improved from giving random outputs to structured sentences not matching the SL sentence, and it eventually learned to produce translation according to the SL sentence.



```
English sentence:
        She had spent so much of her life feeling secondary ( though to a man she adored ) that she burned to undo the laws that hemmed women in . She was only an ordinary woman , she like
d to say .
Correct translation:
        在她一生的大部分时间中，她总感觉自己作为女性得不到重视（虽然她受到一个男人的宠爱），于是尽自己所能解除那些禁锢女性的法律。
Translated as:
        他，，的，的的的，热油烫伤，的的，，，，，的的，，，，，，，，，，，的的的。。
```

*Fig 4.* initial translation without training



```
────────────────────────────────────────────────────
English sentence:
        Even Amazon , an online shopping giant , may feel the urge to splash out and buy , perhaps , Sony-Ericsson , which like Motorola seems too small to make it on
its own .
Correct translation:
        甚至连网购业巨头亚马逊也迫不及待地想进来，它可能会收购索尼爱立信，这家公司与摩托罗拉一样，已经萎缩到无法支撑下去。
Translated as:
        自参议院拉加，的的的，，，，，，，，，，，，飞机，，，，，电锯万份肖像传输放。
────────────────────────────────────────────────────
English sentence:
        So , something at 30 decibels would be pretty faint .
Correct translation:
        因此,30分贝的声音相当微弱。
Translated as:
        乍一看，，是不用费时既的的指头。
────────────────────────────────────────────────────
English sentence:
        This is IN THE NEWS in VOA Special English .
Correct translation:
        这里是美国之音慢速英语时事新闻节目。
Translated as:
        这里是美国之音慢速英语新闻报道。。
────────────────────────────────────────────────────
English sentence:
        This is English as a Second Language Podcast episode 443 .
Correct translation:
        这是ESL博客第443篇。
Translated as:
        我是 你觉得很成功。
────────────────────────────────────────────────────
English sentence:
        IF THERE is any endeavour whose fruits should be freely available , that endeavour is surely publicly financed science .
Correct translation:
        如果有人试图把成果免费公开的话，那他一定是科学资助者，
Translated as:
        当合法买单的的的，授权的的的的年长的的的罪行。
```

*Fig 5.* translation after 3000 epochs

## IV. Conclusion

NMT provides a way for programmers to develop a translation model without specifying the 1-to-1, 1-to-many, or many-to-1 relationships between SL and TL. NMT learned from a great amount of examples about the underlying relationship between SL and TL, and by using sentence embedding techniques, which was analogous to interlingua schema discussed in the first deliverable, the model could produce satisfactory translations.

In the experiment section, it shows that NMT produced valid translations without giving the relationships between SL and TL vocabularies but a lot of sentences. However, it was apparent that the translations were not optimal and had a great room for improvement and the model was experiencing overfitting; this could be solved by increasing the number of training data. Due to lack of computing power, this report will not conduct experiments on those configurations.

# References

[1] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018.

[2] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine,* vol. 29, pp. 141-142, 2012.

[3] I. Sutskever, V. Oriol and Q. V. Le, *Sequence to Sequence Learning with Neural Networks,* arXiv, 2014.

[4] L. Tian, D. F. Wong, L. S. Chao, P. Quaresma, F. Oliveira, Y. Lu, S. Li, Y. Wang and L. Wang, "UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.