Applied Sequence2Sequence Model in Neural Machine Translation

I. INTRODUCTION:

Every language on Earth is used throughout communication, expressing interests or showing emotions of human-beings. It is not hard to learn a foreign language; however, it will be the best if a native speaker can understand every single word of a non-native speaker thoroughly. Linguistics techniques help a lot towards the way a native and non-native speaker communicate. In recent years, linguistics combines with machine learning as well as artificial intelligence techniques has become one of the most interest in terms of neural machine translation.

The motivation of this project is basically based on interest in respect to machine translation and curiosity of the author as to how a neural machine translation model works. This project aims to translate word-by-word, sequence-by-sequence from English to Vietnamese by mainly approaching [1] sequence2sequence neural network model. Following contents describes the basic workflow of this project.

II. DATASET

Dataset is collected from source [2] which includes 3409 sentence pairs as in order English (en) and Vietnamese (vn). The

data text file is named as “en-vn.txt” for preprocessing simplicity.

III. METHODS

a. Preprocessing

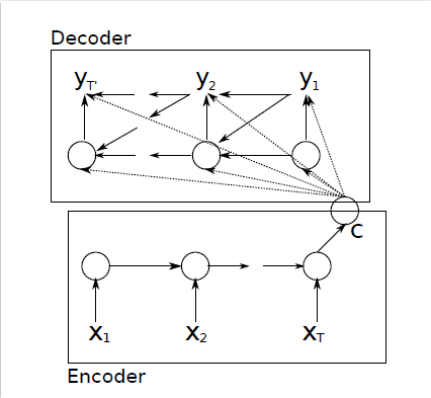
One of difficulties towards the main approach of this project is to preprocess a text sequence from non-tonal language (en) into tonal-language (vn). Firstly, text file is converted from Unicode to ASCII. It helps to remove all tones in Vietnamese which is a way to implement a word as a vector. Furthermore, I set some rules of English prefixes. After preprocessing, the original dataset has 182 sentence pairs left which included 292 English words and 290 Vietnamese words.

The processing step pretty much explains the big difference between non-tone language and tonal-language.

b. Sequence2Sequence model

1. A model network consists of two Recurrent Neural Networks (RNNs) called encoder and decoder.
2. Encoder: the first RNN reads the input sequence (en) to output a single vector.

3. Decoder: the second RNN which reads the output vector



from encoder to produce the output sequence (vn).

Figure 1. Encoder-Decoder model. [3]

IV. EXPERIMENT

Throughout implementation, tokens start-of-sequence <SOS> and end-of-sequence <EOS> are applied as to gain the accuracy when normalizing each word of a sequence with separator to vector.

To overcome the vanishing gradient problem in a traditional RNN network, gated recurrence unit (GRU) is also applied to solve the problem. It is a gated mechanism which basically helps to figure out what information is passed into the output of the model.

During training phase, epoch is set to 25000 with 256 hidden states.

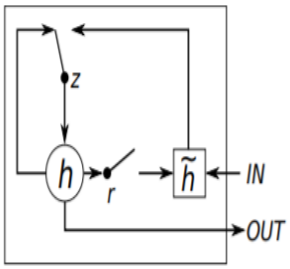


Figure 2. Gated recurrence unit diagram. [4]

V. RESULT

The result comes out as expected which somehow translates English sequences into Vietnamese sequences (without tones version).

```
0m 55s (- 3m 41s) (5000 20%) 1.4021
1m 53s (- 2m 50s) (10000 40%) 0.0502
2m 51s (- 1m 54s) (15000 60%) 0.0130
3m 46s (- 0m 56s) (20000 80%) 0.0073
4m 43s (- 0m 0s) (25000 100%) 0.0050
```

Figure 3. Total loss result.

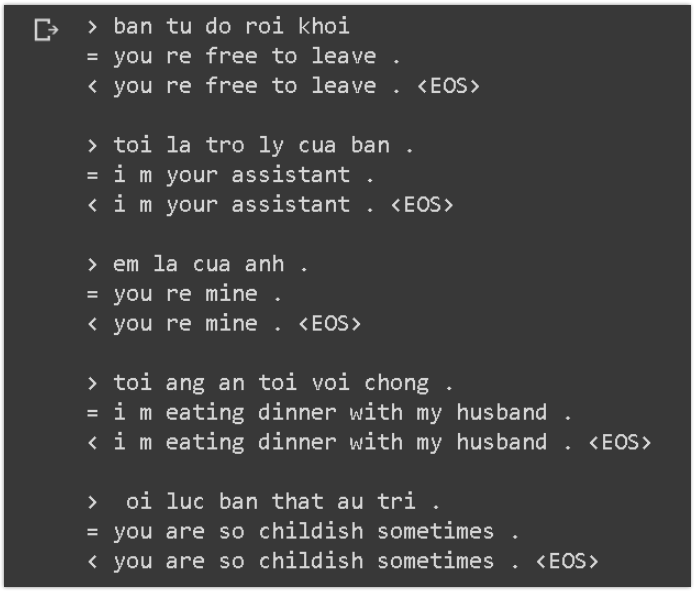


Figure 4. English sentences to Vietnamese sentences.

VI. DISCUSSION

The result gives a decent accuracy. However, some letter which is not in English alphabets cannot be translated since we already converted all letters to ASCII standard from the beginning. It is seen to be one of motivations for the future work in which we can try apply POS tagging technique.

VII. CONCLUSION

The project has successfully implemented sequence2sequence network model as approaching to the main goal. The model gives us an acceptable outcome regardless to our English- Vietnamese translating model. Besides its pros, there is a big disadvantage in which brings out the inefficiency of computer memory when we have a big dataset.

VIII. REFERENCES

[1] I. Sutskever, O. Vinyals and Quoc Le, "Sequence to Sequence Learning with Neural Networks," 2014.

[2] tatoeba.org

[3] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Leanring Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation," 2014.

[4] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 2014.

[5] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," ICLR 2015.

[6] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Leanring Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation,"

[7] S. Bird, E. Klein and E. Loper, “Natural Language Processing with Python,” O’reilly.

[8] D. Jurafsky and J. Martin, “Speech and Language Processing,” 3rd ed. Draft, 2018.