English to Chinese Neural Machine Translation

CSCI 3907/6907

Zhibo Sheng, Nopphiphat Suraminitkul, Christopher White

Introduction

In this project, our team developed a neural machine translation model that is capable of translating sentences from English to Chinese. Our team designed and implemented a sequence to sequence neural machine translation model (NMT) using Python and TensorFlow. We used the News Commentary corpus to train our model and evaluate its performance using an alternative to the Bi Lingual Evaluation Understudy (BLEU) metric, called the Rank-based Intuitive Bilingual Evaluation Score (RIBES), since it is designed to evaluate Asian languages [1] [2]. This paper discusses the background of NMT, the approach for our model, and the results produced by our novel model for English to Chinese translation.

As the world becomes more connected through the internet, the demand for automated translation between different languages is rapidly growing. Language translation overcomes language barriers, bridges people from one part of the world to another, expands business consumer base internationally, promotes the effective exchange of ideas and works, and more. The history of automated translation began with rule-based machine translation (RBMT) in the 1950s including direct machine translation, transfer-based machine translation, interlingual machine translation, and statistical machine translation (SMT). NMT, proposed by Kalchbrenner and Blunsom (2013), Sutskever et al. (2014), and Cho et al. (2014), is a newly emerging approach to machine translation that has risen in popularity in the past 5 years and is demonstrating remarkable performance in a variety of application areas [3][4][5]. Unlike traditional translation systems, which consists of sub-components that are tuned separately, the neural network approach utilizes deep learning neural networks that reads a sentence, maps it among natural language, and outputs a translation.

Most NMT models consist of sequence-to-sequence (seq2seq) modeling using recurrent neural networks (RNN) with an encoder and a decoder architecture. The encoder takes an input sentence and encodes it into a fixed-length vector. The decoder then produces a translated output from the encoded context vector. This encoder-decoder system, which comprised of several RNN layers, is jointly trained to maximize the probability of a correct translation given a source sentence.

Related Work

Machine translation, or the approach that utilizes a computer to do “automatic translation of text from one natural language (the source language) to another (the target language)” [9], is a highly active area of research with many published studies in the past several years. The initial challenge of “long distance reordering” for the first NMT model was later addressed with the introduction of RNN [10], and shifted the attention towards the problem of the “exploding/vanishing gradient” [11], which makes RNN hard to actually handle long distance dependencies as proven by the rapid deteriorating performance of the basic encoder-decoder model with increasing input sentence length [5]. It was later with the application of Long Short-Term Memory (LSTM) (2014) and the “attention” mechanism (2015) that the problems of “exploding/vanishing gradients” and “fixed-length vector” are solved.

Several recent studies have addressed the challenging process of translation of text between English and Chinese using neural networks. Innovative enhancements to neural network based algorithms include synchronous bidirectional decoding in a single model [6]. Additionally, LSTM models built with encoder and decoder architectures with attention mechanisms have been implemented to improve translation accuracy [7]. Another novel approach involves post-editing process of the recently-introduced neural machine translation [7] [8]. In this paper, we will explore the encoder and decoder LSTM architecture with bidirectional LSTMs in the encoder with an attention layer. The attention layer to some extent allows the neural network to correlate words in a sentence.

Approach

The model was trained and evaluated using the News Commentary corpus, which includes political and economic commentary pulled from articles on Project Syndicate, obtained from the Second Conference On Machine Translation. It contains more than 300,000 English sentences, which were translated into Chinese using the TensorFlow wmt17_translate function which in order to generate the parallel corpus. Once the parallel corpus was created, we removed language specific punctuation for both English and Chinese sentences. Next, we used NLTK to tokenize the English sentences and used Jieba to tokenize the Chinese sentences, since it is challenging to tokenize Chinese words based on whitespace or characters alone. Once the sentences were tokenized, the sequence length for all sentences was fixed at a maximum length of 10 words per sentence due to memory limitations. Finally, we added ‘<start>’ and ‘<end>’ tokens to the beginning and end of the cleansed English and Chinese sentences in preparation for the model training.

We then split the News Commentary corpus into 2 datasets with 99.84% of the data used for the teacher forcing method used for training our model and the remaining 0.16% (approximately 500 samples) for testing to evaluate the model’s performance. In order to represent the words in a format that machine learning algorithms can understand, we represent words as frequency index vectors.

We used an iterative approach to experiment with various neural network architectures and natural language processing techniques. Our initial NMT model consisted of GRU layers within encoder and decoder blocks. For the final model we introduced a bidirectional LSTM layer in the encoder which originally contained an embedding layer and GRU. The final hidden states of the bidirectional LSTM are averaged before the output is passed to the decoder and attention layer. Next is the decoder which consists of an attention layer, in this case the Bahdanau Attention layer which calculates alignment and outputs the context vectors which are concatenated with the input embedding layer for the decoder. This result is passed through a GRU which passes through a fully connected layer and returns the predicted translation sequence.

The neural network is trained using teacher forcing which is a strategy for training neural networks. It uses the model output from a prior timestep as input. It uses the ground truth output from the training dataset at the current time step as input for the next timestep rather than using the output generated by the network. The start and end tokens at the beginning and end of each sequence signal the start and end of the sequence. Teacher forcing is a quick and effective way to train recurrent neural networks.

We used both qualitative and quantitative measures to evaluate the performance using RIBES evaluation metric. It is a metric that is used to compare the original text and the translated output of the neural network developed by NTT Communication Science Labs [5]. Additionally our group consists of English and Mandarin speakers, so we will be able to perform qualitative evaluation of translations.

Experimental Setup

The final NMT model consisted of 1024 hidden units, a batch size of 128, 20 training epochs, and 256 dimensional embeddings. We used an Azure cloud computing platform for full scale training and evaluation of our model. The instance contains a 6 core CPU, 56 Gb of memory, and a K80 GPU. We used Python to develop our model and utilized libraries such as NLTK and Jieba to preprocess our data, and we use TensorFlow 2.0 to build our model.

Results and Analysis

The final model trained using 20 epochs produced an average RIBES score of 0.2241 for the 500 test sentences. Some sample predicted translations are are displayed in Tables 1 and 2 along with the corresponding English and Chinese ground truth. Overall the model performed well with the model producing some perfect translations of the test sentences

English	Chinese Ground Truth	Chinese Predicted	RIBES
['Is Economics a Science']	['经济学', '是', '科学', '吗']	['经济学', '是', '科学', '吗']	1
['Globalization and its New Discontents']	['全球化', '及其', '新', '的', '不满情绪']	['全球化', '及其', '新', '的', '不满情绪']	1

Table 1: Examples of Correct Testing Set Translations

English	Chinese Ground Truth	Chinese Predicted	RIBES
['Asias choice is clear : either the region embraces a']	['亚洲', '所', '面临', '的', '选择', '是', '一清二楚', '的', '要么', '就是']	['亚洲', '的', '选择', '要么', '是', '明确', '的']	0.0881
['Another would be to create new visa categories to enable']	['另', '一个', '办法', '是', '建立', '新', '的', '签证', '种类', '让']	['另', '一种', '新', '的', '是', '新', '签证', '申请']	0.0867

Table 2: Examples of Incorrect Testing Set Translations

Conclusion and Project Insights

The bidirectional NMT with attention layer produced reasonably accurate results for the testing sentences. This project topic was interesting because translation between English and Chinese is challenging due to the fundamental differences in structure and meaning. Initially, the model was producing meaningless results due to several problems that were addressed based on the presentation feedback. The first update was increasing the training data from 30,000 to approximately 300,000 records, which was facilitated by limiting the length of the sentences to 10 words. Additionally, further preprocessing was conducted to remove the Chinese punctuation, which is different from the punctuation used in English. Furthermore, the introduction of the bidirectional LSTM allowed the model to gain forward and backward sentence context, which allowed it to generate more accurate translations. Finally, we overcame the low BLEU score we reported during the presentation by switching to the RIBES metric, which is specialized for evaluating Chinese and Japanese sentences.

Future Work

While our model is capable of producing accurate English to Chinese translations, there are still further improvements that can be made to increase the performance of the algorithm. Deep learning models can always be adjusted, one of these methods is hyperparameter tuning, which consists of adjusting the hidden unit size within layers, changing layer types, or experimenting with different attention mechanisms. Due to time constraints, only 20 epochs were feasible to execute and therefore additional training epochs can further strengthen neural network weights which will allow for more accurate translations. Additionally, increasing sentence length to 15 - 20 words would allow the model to better capture the context of English sentences.

References

[1] K. Papineni, “BLEU: a Method for Automatic Evaluation of Machine Translation,” *Proceedings of the 40th annual meeting on association for computational linguistics.*, 2002.

[2] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, “Automatic Evaluation of Translation Quality for Distant Language Pairs,” *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Oct. 2010.

[3] N. Kalchbrenner and P. Blunsom, “Recurrent continuous translation models,” *Kalchbrenner, Nal, and Phil Blunsom. “Recurrent continuous translation models.” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.*, 2013.

[4] I. Sutskever, O. Vinyals, and Q. V. Le., “Sequence to Sequence Learning with Neural Networks,” *Advances in NIPS*, 2014.

[5] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[6] L. Zhou, J. Zhang, and C. Zong, “Synchronous Bidirectional Neural Machine Translation,” *Transactions of the Association for Computational Linguistics*, vol. 7. pp. 91–105, 2019.

[7] Y. Wang *et al.*, “Sogou Neural Machine Translation Systems for WMT17,” *Proceedings of the Second Conference on Machine Translation*. 2017.

[8] Y. Jia, M. Carl, and X. Wang, “Post-editing neural machine translation versus phrase-based machine translation for English–Chinese,” *Machine Translation*, vol. 33, no. 1–2. pp. 9–29, 2019.

[9] S. Russell and P. Norvig, “Artificial intelligence: a modern approach”, 1995.

[10] K. Sudoh, K. Duh, H. Tsukada, T. Hirao, and M. Nagata, “Divide and translate: improving long distance reordering in statistical machine translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (pp. 418-427). Association for Computational Linguistics, 2010.

[11] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks. In International Conference on Machine Learning, pp. 1310-1318, 2013.