A Comparative Study of Seq2Seq and Transformer Models for Natural Language Processing Tasks

Hwang Gyubin

September 21, 2023

Abstract

This study utilized two deep learning models, Seq2Seq and Transformer, to compare the performance of the Korean-English translator model. We analyze the differences in translation results from the two models and demonstrate that a particular model performs well based on the experimental results. These results suggest that we can expect better performance when performing translation tasks using the corresponding model and suggest that it is an important determinant in the model selection process for natural language processing tasks.

1 Introduction

Recent developments in natural language processing have brought a new leap forward in machine translation. In particular, Korean-English translation work is increasing in importance in the field of international communication and business, and innovative changes are taking place in this field due to the development of deep learning technology.

The main objective of this study is to compare the performance of two key deep learning models, Seq2Seq and Transformer, in performing Korean-English translation, and to demonstrate that Transformer models perform well. Through this comparison, we would like to demonstrate experimentally in our paper that the Transformer model outperforms the Seq2Seq model in translation tasks.

This paper emphasizes the need for research, emphasizes the importance of modern translation technology, and compares the pros and cons of Seq2Seq and Transformer models to argue that the Transformer model is better suited for translation work. In addition, we experimentally verify this argument and provide guidelines for model selection in natural language processing based on its findings. The structure of the paper is as follows.

2 Background and Related Works

2.1 Seq2Seq model

The Seq2Seq model (Sequence-to-Sequence) is a model used specifically for machine translation in natural language processing tasks. The model works by utilizing the Recurrent Neural Network (RNN) family of networks to encode the input sequence into a vector of a fixed length and then re-decode the vector to generate an output sequence. The Seq2Seq model initially performed well in machine translation and improved translation quality with the introduction of the Attachment mechanism.

2.2 Transformer model

The Transformer model was developed to overcome the limitations of the Seq2Seq model, which brought about a revolutionary change in natural language processing. The Transformer model is based on the Attachment mechanism, which utilizes Self-Attention instead of RNN to model the relationship between input and output sequences. This model is easy for parallel processing, has strength in long sequence processing, and performs well in machine translation and natural language understanding tasks.

2.3 related research

Research related to the comparison of Seq2Seq and Transformer models is actively being conducted in the field of natural language processing. Previous studies have performed an analysis of the performance differences and features of the two models, and it has been reported that the Transformer model has performed well in several natural language processing tasks. However, under certain tasks and conditions, the Seq2Seq model may still be useful, and studies on selection criteria in these situations are also ongoing. These findings will help us better understand the pros and cons of the two models.

3 Method

3.1 Dataset

The study uses a parallel Korean-English dataset called "Korean Parallel Corporation." This dataset includes Korean sentences required for translation work and English translation of those sentences, and covers high quality and diverse topics. It is divided into learning data and evaluation data for experiments.

3.2 Model architecture

3.2.1 Seq2Seq Model

The Seq2Seq model consists of an encoder and a decoder. The encoder encodes the input sentence into a vector of fixed length, and the decoder generates the translation sentence using the output from the encoder and the output from the previous step. This model is widely used in natural language processing tasks and uses a two-way LSTM (Bi-LSTM) to contextualize.

3.2.2 Transformer model

The Transformer model consists of multiple layers each of the encoder and decoder. Each layer consists of multi-head attention and feedforward neural networks, which leverage the Self-Attention mechanism to model the relationship between input sequences. Positional Encoding adds word location information and has advantages in parallel processing and long sequence processing.

3.3 Model Train

All experiments pre-learn each model with a "Korean Parallel Corpora" dataset. However, data preprocessing and tokenization are required before model learning. This process is carried out as follows:

- Refines or normalizes unnecessary strings such as special characters, spaces, and punctuation.
- For the Seq2Seq model, Korean sentences are divided into morpheme units using Mecab, a morpheme analyzer, and sequences are prepared by adding a start token and an end token.
- For the Transformer model, Korean sentences are divided into token units using a tokenization library such as Sentence Piece, and model inputs are constructed by adding a start token and an end token.

In the course of learning, we use the cross entropy loss function to minimize loss and apply gradient clipping to prevent gradient congestion. All learning is efficiently performed utilizing GPU acceleration, and each model is trained for a given number of epoches.

Prepare the learning data for the model to understand and learn the learning data effectively through the data preprocessing and tokenization process above.

3.4 Model Evaluation

The performance of the model is measured by the Buildingual Evaluation Understudy (BLEU) score. This score is used to measure the similarity between the model-generated translation results and the human translation results. Higher BLEU scores indicate better translation quality. Use the same metrics to compare the performance of Seq2Seq and Transformer models.

4 Result

In this study, we performed Korean-English translation work using Seq2Seq and Transformer models. Each model was trained using the ”Korean Parallel Corpora” dataset and used Building Evaluation Understudy (BLEU) scores to measure translation performance.

Experiments showed that the Transformer model outperformed the Seq2Seq model. The Transformer model leverages multi-head attention and positional encoding to model the relationship between input sequences, resulting in high translation quality. However, the experiment is currently underway, and visualization data and additional results will be collected and analyzed. We will update the paper afterwards.

5 Acknowledgment

The author is deeply grateful to his colleagues for their support in conducting this study and writing the paper. In particular, I would like to express my deep appreciation for the support and cooperation at everyone's lab Aiffel. This research was possible with the support and guidance of everyone's lab Aiffel, which improved the quality and successful completion of the research. I also thank my research colleagues for giving me advice and feedback during the research process. We would also like to thank everyone's laboratory for providing datasets and research equipment.

This study was not possible without support and cooperation from everyone's lab Aiffel, and I am deeply grateful for the support so far.

References

[1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

[2] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint* arXiv:1409.0473.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 30-38).

[4] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint* arXiv:1609.08144.

[5] Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint* arXiv:1409.1259.

[6] Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., ... & Zaremba, W. (2018). Tensor2tensor for neural machine translation. *arXiv preprint* arXiv:1803.07416.