Differences in Translation Tendencies Between Seq2Seq and Transformer Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study evaluates and compares the translation tendencies of Seq2Seq and Transformer models. We analyzed the translation results using a dataset consisting of sentence pairs in Korean and English. The Seq2Seq model, which utilizes Bahdanau attention and GRU layers, was compared against a Transformer model with parameters as described in the original paper [1]. Both models were trained for 10 epochs with a consistent batch training time, ensuring a fair comparison. Our analysis revealed that the Seq2Seq model tends to produce translations based on word-to-word matching, which can lead to grammatical errors or loss of context if the model fails to find the correct translation. In contrast, while the Transformer model often generates more natural-sounding translations, it sometimes distorts the context or produces irrelevant output. These findings highlight the strengths and limitations of each model type, providing insights for future improvements in machine translation systems.

1   Introduction

Advancements in machine translation have led to the development of various models. Among these, Seq2Seq with attention and transformer models have shown significant promise. I intend to compare the translation tendencies of two models: Seq2Seq with attention and transformer. To make a fair comparison, I will set both models to a similar scale and train them using the same data. Additionally, to evaluate the results, I will provide the same sentences to both models and analyze the outputs qualitatively.

2   Method

2.1   Train Data

The data is *korean-english-news-v1*, subset of *korean-parallel-corpora*[1]. It consists of about 75K sentence pairs from various news article in Korean and English, not entire articles.

The word token vocabulary size is 20k. According to the transformer paper [1], they set the vocabulary size to around 30k. However, since my data is smaller, I decided to set the size to 20k. Additionally, the word embedding size is 512, which is the same as in the paper.

---

[1]https://github.com/jungyeul/korean-parallel-corpora/tree/master/korean-english-news-v1

Table 1: Translation result

No	Original	Translation	
		Seq2Seq	transformer
1	오바마는 대통령이다	obama is the president	president obama is in his country
2	시민들은 도시 속에 산다	people in the city	city animals are hoern soaring
3	커피는 필요 없다	it s not a caffeine	it needs
4	일곱 명의 사망자가 발생했다	seven people were killed	seven of the dead were civilians

2.2 Models

To compare different model architectures, the batch training times are adjusted to be similar. Using an Nvidia 4070 Ti, the time for one epoch is approximately 17 minutes. Both models are trained for 10 epochs each.

2.2.1 Seq2Seq with attention

The first model is Seq2Seq with attention. Specifically, it uses Bahdanau Attention and has GRU layers with layers size of 1024.

2.2.2 Transformer

The second model is a transformer with parameters matching those in the paper [1]. It uses positional encoding, has 6 layers in both the encoder and decoder, 8 attention heads, and a feed-forward network size of 2048.

3 Result

The results of the translation are shown in Table 1.

3.1 Tendency of Seq2Seq

Seq2Seq tends to produce outputs based on word-to-word matching, like finding a translated word using a dictionary and replacing that word grammatically. If Seq2Seq makes a mistake (i.e., fails to translate correctly), it may select an incorrect word from the dictionary, distorting the context of the original sentence, or fail to find a suitable word, thus missing the context of the original sentence. Also it tends to maintain structure, but can still make grammatical errors. For example, as shown in the table, Seq2Seq has difficulty distinguishing between count nouns and non-count nouns.

3.1.1 Qualitative analysis

No 1 : Significantly accurate. But 'President Obama' would be a more natural expression.
No 2 : Missing a verb. While 'people in the city' conveys a similar meaning, it weakens the context of 'live (산다)'. 'people live in the city' is more natural.
No 3 : Missing a verb. 'it (doe)s not (need) a caffeine' would be correct.
No 4 : Accurate.

3.2 Tendency of transformer

The Transformer model tends to produce outputs that are more natural and context-rich. Compared to Seq2Seq, these translations are more narrative and grammatically fluent. While the Transformer has an abundant expression capacity and higher grammatical accuracy, it sometimes generates sentences with additional context that is not present in the original text. For example, in No 4, the translations from other epochs often include information about the day of the week.

- Ex 1: "Earlier Sunday, eight people died in the death toll."
- Ex 2: "On Wednesday, the second death toll from the dead."

3.2.1 Qualitative analysis

No 1 : Includes additional context. The location where he is cannot be specified.
No 2 : Completely wrong. The model generates a sentence based solely on the word 'city.'
No 3 : Completely wrong. The model only understands the word 'need (필요).'
No 4 : Includes additional context. While the dead may be civilians, this is questionable.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. Neural Information Processing Systems (NeurIPS). `https://arxiv.org/abs/1706.03762`