Analysis of Alignment Collapse and Overfitting in Bahdanau Attention-based Seq2Seq Models under Low-Resource Conditions

Wonjin Choi
Dankook University
Yongin, Republic of Korea
cwj2238@naver.com

Abstract

While Neural Machine Translation (NMT) outperforms statistical methods, it inherently requires extensive parallel corpora to achieve generalization. This study investigates the training instability and performance degradation of a Seq2Seq model equipped with the Bahdanau Attention mechanism under extreme low-resource conditions, utilizing a dataset of approximately 6,000 sentence pairs. To mitigate data sparsity, we employed the SentencePiece tokenizer and experimentally restricted the vocabulary size to 2,000. The results demonstrate that while the model showed a rapid decrease in validation loss during the initial training phase, it quickly succumbed to overfitting, exhibiting severe repetition errors and context loss in the generation phase[1]. Crucially, through the visualization analysis of Attention Maps, we identified that data scarcity leads directly to a failure in learning the alignment between the encoder and decoder. The attention weights failed to focus on specific source tokens, appearing blurred or exhibiting a tendency to rely solely on the decoder's language modeling priors rather than the source context. These findings suggest that in the absence of sufficient data, complex attention mechanisms may act as noise. Consequently, this study verifies that transfer learning from pre-trained models or robust data augmentation techniques is essential for NMT in low-resource settings.

1 Introduction

With the rapid advancement of deep learning, the field of Machine Translation has undergone a complete paradigm shift from Statistical Machine Translation (SMT) to Neural Machine Translation (NMT). In particular, the Sequence-to-Sequence (Seq2Seq) model, which compresses an input sequence into a fixed-length vector to generate an output sequence, has established itself as a standard architecture not only for machine translation but also for various natural language processing tasks such as text summarization and chatbots.

Early Seq2Seq models suffered from a bottleneck problem where information loss occurred as the input sentence length increased, due to the compression of all information into a fixed-size context vector. To address this, Bahdanau et al. [1] proposed the Attention Mechanism, which allows the decoder to refer back to the encoder's input sequence at each time step, dynamically assigning weights to relevant parts. This innovation significantly improved translation performance by preserving source information.

However, the success of NMT models is heavily predicated on the availability of large-scale parallel corpora. Koehn and Knowles [2] pointed out that in low-resource environments where data is scarce, NMT models perform significantly worse than statistical methods. They specifically highlighted

NeurIPS 2021 AI for Science Workshop.

vulnerabilities such as hallucinations, where the model ignores context in favor of fluency, and the infinite repetition of specific words.

This study initiates from this problem statement and provides an in-depth analysis of the training collapse experienced by the Bahdanau Attention model under an extreme low-resource environment consisting of approximately 6,000 sentence pairs. To mitigate data sparsity, we applied the SentencePiece tokenizer and set an experimental constraint limiting the vocabulary size to 2,000. The primary contribution of this paper goes beyond merely measuring translation performance (BLEU); we aim to structurally identify the impact of data scarcity on the failure of alignment learning through the visualization of Attention Maps.

2 Related Work

2.1 Seq2Seq & Attention

The Sequence-to-Sequence (Seq2Seq) model, which has become the standard in machine translation, features an end-to-end architecture consisting of an encoder and a decoder. The encoder compresses the input sequence into a fixed-dimensional context vector, from which the decoder generates the target sequence. However, this architecture suffers from a bottleneck problem where information loss occurs as the input sentence length increases, making it difficult to encapsulate all semantic information into a fixed-size vector.

To address this limitation, Bahdanau et al. [1] proposed the Attention Mechanism. Attention allows the decoder to refer back to all hidden states of the encoder at each time step (t), dynamically assigning weights (Alignment Scores) to the parts highly relevant to the word currently being predicted. Formally, the context vector c_t at decoder time step t is calculated as a weighted sum of the encoder hidden states h_j and attention weights α_{tj}:

$$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j \tag{1}$$

Through this mechanism, the model can effectively preserve source information even in long sentences and self-learn the alignment between source and target languages without explicit supervision. In this study, we adopt this architecture to analyze its behavior under constrained data environments.

2.2 Challenges in Low-Resource NMT

Despite the significant performance improvements brought by the attention mechanism, Neural Machine Translation (NMT) inherently relies on large-scale parallel corpora containing millions of sentence pairs. In "low-resource" settings where data is insufficient, NMT models are highly susceptible to overfitting, a common issue in deep learning architectures.

Koehn and Knowles [2] demonstrated that NMT performance degrades sharply compared to traditional Statistical Machine Translation (SMT) in low-resource environments, out-of-domain scenarios, and rare word handling. Specifically, they pointed out that under data scarcity, models tend to ignore source context and prioritize fluency, leading to issues such as the infinite repetition of specific words or hallucinations, where the output is completely unrelated to the input. This occurs because the model fails to learn precise alignment between source and target sentences and relies excessively on the decoder's language model priors. Based on this theoretical background, this study aims to visually identify the "alignment collapse" phenomenon in attention maps when trained with an extremely small dataset of approximately 6,000 pairs.

3 Methodology

3.1 Dataset and Preprocessing

In this study, we utilized a parallel corpus consisting of approximately 6,000 sentence pairs for the Korean-English translation task. This is an extremely small amount compared to standard NMT datasets, serving as an intentional constraint to analyze model behavior in low-resource environments. For preprocessing, we applied normalization to remove noise, followed by tokenization using Google's SentencePiece.

Specifically, considering the data sparsity, we strictly limited the vocabulary size to 2,000. This is significantly smaller than the typical range of 8k to 32k, aimed at evaluating how efficiently the model can compress context and align sequences with limited expressiveness. Start tokens <start> and end tokens <end> were appended to all sentences, and padding was applied with a maximum sequence length of 40 to facilitate efficient batch processing.

3.2 Model Architecture

We adopted a Seq2Seq model based on Bahdanau Attention [1]. The architecture consists of an Encoder, an Attention Layer, and a Decoder(Figure 1).

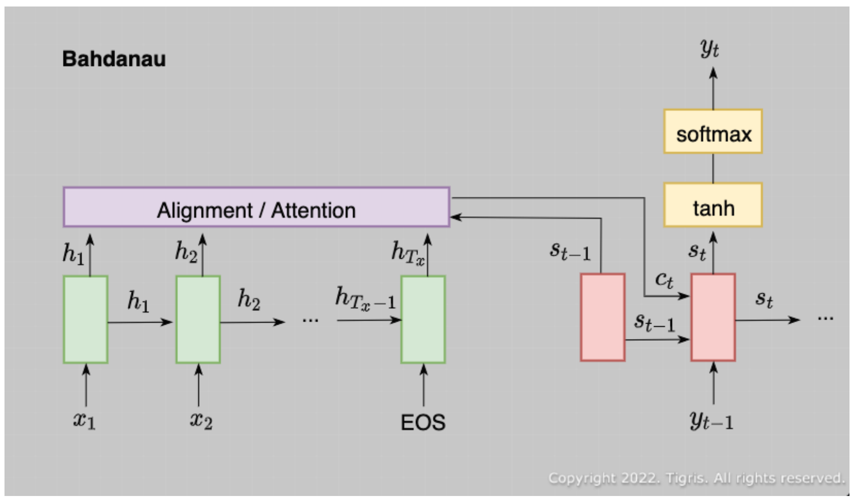

Figure 1: Seq2Seq model based on Bahdanau Attention

3.2.1 Encoder

The input token sequence passes through a 128-dimensional embedding layer and is processed by a Gated Recurrent Unit (GRU) with 256 units. The encoder hidden state h_t at each time step t encodes the information of the input sentence.

3.2.2 Attention Mechanism

When the decoder predicts a word at time step t, it calculates alignment scores for all encoder hidden states h_s. In this study, we employed the Concat (Additive) score function defined as:

$$\text{score}(h_t, h_s) = v_a^T \tanh(W_a h_t + U_a h_s) \tag{2}$$

where W_a, U_a, v_a are trainable parameters. The calculated scores are normalized via a Softmax function to obtain attention weights α_{ts}, which are then used to generate the context vector c_t.

3.2.3 Decoder

The decoder receives the context vector c_t and the output word from the previous time step, processes them through a 256-unit GRU, and finally outputs the probability distribution of the next word via a Fully Connected Layer.

4 Experimental Setup

4.1 Implementation and Training Details

This experiment was implemented using the TensorFlow framework. We utilized the Adam optimizer for model training and adopted Sparse Categorical Crossentropy as the loss function to efficiently handle integer-encoded target sequences. The hyperparameters were fixed with a batch size of 32 and a maximum of 50 epochs, with the vocabulary size restricted to 2,000 as described in Section 3.1.

Crucially, to mitigate the exposure bias inherent in Seq2Seq models and ensure training stability in a low-resource setting, we implemented a Scheduled Sampling (Dynamic Teacher Forcing) strategy. In the initial phase, the teacher forcing ratio was set to 1.0 (feeding ground truth) to facilitate rapid convergence. As training progressed, this ratio was linearly decayed, gradually forcing the model to rely on its own predictions. The teacher forcing ratio ϵ_e at epoch e is defined as:

$$\epsilon_e = \max\left(0.0, 1.0 - \frac{e}{K}\right) \tag{3}$$

where K represents the decay epochs. In our experiment, we set $K = 20$, meaning that after 20 epochs, the ratio becomes 0.0, and the model performs inference completely independently. Furthermore, to prevent overfitting, Early Stopping (Patience=10) was applied, terminating the training process if the validation loss did not improve for 10 consecutive epochs.

4.2 Evaluation Metrics and Analysis Methods

Given the experimental nature of using extremely limited data and vocabulary, this study focuses on analyzing the structural behavior of the model rather than relying solely on standard quantitative metrics like BLEU scores.

4.2.1 Learning Curve Analysis

We compare the trends of Training Loss and Validation Loss to identify the onset of overfitting and verify model convergence.

4.2.2 Attention Map Visualization

We visualize the attention weights as heatmaps to qualitatively evaluate the success of alignment learning. A well-trained model should exhibit a clear diagonal pattern, indicating that the decoder attends to the corresponding source tokens.

4.2.3 Error Taxonomy

We categorize major errors in the generated translations into types such as 'Repetition', 'Context Loss', and 'Out-Of-Vocabulary (OOV)' to provide an in-depth analysis of the impact of the low-resource environment on model performance.

5 Results & Analysis

5.1 Training Dynamics & Overfitting

Training the model with approximately 6,000 sentence pairs over 50 epochs revealed a distinct divergence between Training Loss and Validation Loss. In the initial 10 epochs, both loss values decreased rapidly, indicating the model was quickly learning data patterns. However, around epoch 15, while the training loss continued to decline to approximately 0.8, the validation loss stagnated around 4.14 or showed a slight increase(Figure 2).

This is a classic symptom of overfitting, suggesting that instead of learning general translation rules between the source and target languages, the model optimized itself by merely memorizing specific sentence patterns within the training data. In particular, the restricted vocabulary size of 2,000 and the small dataset size limited the model's ability to capture subtle linguistic nuances. Despite the training being terminated early by Early Stopping, the model failed to achieve generalized performance.

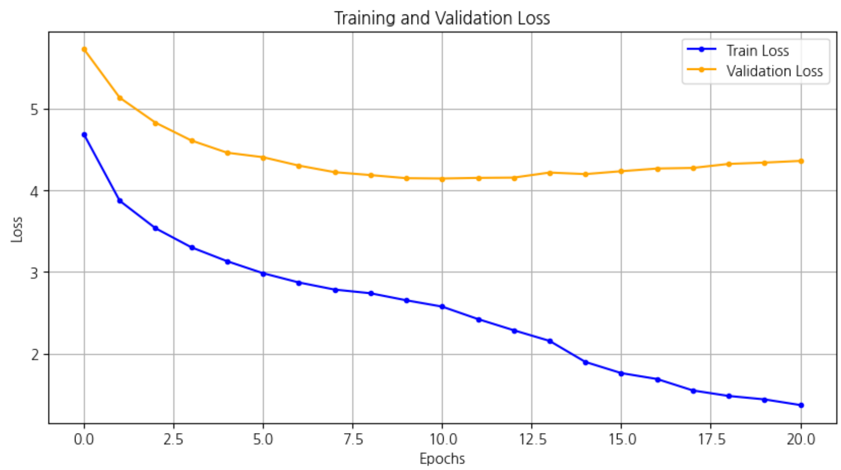

Figure 2: Training and Validation Loss

5.2 Qualitative Evaluation of Translation

To complement the limitations of quantitative metrics, we qualitatively analyzed the actual translations generated by the trained model. [Table 1] illustrates major failure cases. The most prominent error type was 'Repetition'. For instance, given the input "Can I have some coffee?", the model generated " 커피 좀 좀도도도 돼 ?" (meaningless repetition of tokens).

Table 1: Translation Examples

Source (English)	Target (Korean)	Prediction	Error Type
may i help you ?	무엇을 도와드릴까요?	내가까 ?까 ?	Repetition & Grammar Collapse
can i have some coffee ?	커피를 좀 주시겠어요?	커피 좀 좀도도도 돼 ?	Repetition
how many apples are there ?	거기 사과가 몇 개 있나요?	거기 사과 몇 얼마나 ?	Partial Success (Memorization)

This phenomenon is attributed to Language Model Bias, where the decoder relies excessively on its own previous outputs rather than the context vector from the encoder when determining the next word. Furthermore, vulnerabilities typical of low-resource NMT as pointed out by Koehn and Knowles [2] were reproduced, such as the omission of key semantic components like verbs or subjects, and the generation of grammatically incoherent sequences.

5.3 Alignment Collapse Analysis via Attention Maps

The visualization of Attention Maps, the core analysis of this study, clearly reveals the fundamental cause of performance degradation. In an ideal Seq2Seq model, a diagonal alignment pattern should emerge, showing high weights at position (i,j) when the i-th word of the source sentence is translated into the j-th word of the target sentence [1].

However, as shown in [Figure 3], the attention maps in our experiment exhibited 'Alignment Collapse', where weights were either blurred across the sequence without focusing on specific source tokens or fixed at certain positions regardless of the decoding step. This indicates that the attention mechanism failed to learn which parts of the input sentence to attend to due to data scarcity. Consequently, the decoder received ambiguous context vectors, which directly led to the aforementioned repetition errors and hallucinations.

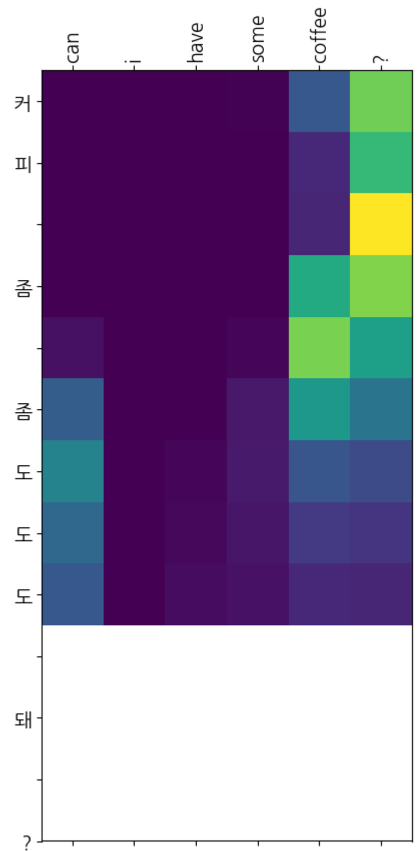

Figure 3: Attention Map of "can i have some coffee ?"

6 Conclusion

In this study, we implemented a Seq2Seq model with Bahdanau Attention under an extreme low-resource environment consisting of approximately 6,000 sentence pairs and provided an in-depth analysis of its structural limitations and causes of failure. Despite applying the SentencePiece tokenizer and a reduced vocabulary size of 2,000 to mitigate data sparsity, the model exhibited rapid overfitting to the training data.

Crucially, the qualitative evaluation and visualization of attention maps clearly demonstrated the phenomenon of 'Alignment Collapse'. The model failed to effectively transfer context information from the source sentence to the decoding phase, which consequently led to infinite repetition of specific tokens and context-irrelevant hallucinations. The results of this experiment reaffirm that attention mechanisms require a certain threshold of data scale to function effectively.

Therefore, for future work, we propose two directions to overcome the limitations of low-resource settings. First, Data Augmentation techniques (e.g., Back-Translation) should be applied to enhance the diversity of training data [3]. Second, it is necessary to move beyond the limitations of RNN architectures by adopting the Transformer model [4], which is advantageous for parallel processing, or by leveraging Transfer Learning from large-scale Pre-trained Language Models (PLMs).

References

[1] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR 2015*.

[2] Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28–39.

[3] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. *ACL 2016*.

[4] Vaswani, A., et al. (2017). Attention Is All You Need. *NeurIPS 2017*.