The Experiment Report of *Machine Learning*

School: School of Software Engineering

Subject: Software Engineering

Author:
Yuming Jiang

Supervisor:
Mingkui Tan

Student ID:
202330550601

Grade:
Software Class 3

November 21, 2025

Sequence-to-Sequence Neural Machine Translation with Attention Mechanism

***Abstract*—This experiment implements a Sequence-to-Sequence (Seq2Seq) neural machine translation system with attention mechanism for Chinese-to-English translation. The model consists of a GRU-based encoder and an attention-enhanced GRU decoder. We train the model on a filtered dataset of 459 sentence pairs from the Tatoeba corpus, achieving a BLEU score of 0.9718 on the evaluation set. The experiment demonstrates the effectiveness of the attention mechanism in capturing long-range dependencies and generating accurate translations. We analyze the attention weights visualization to understand the model's alignment behavior between source and target sequences.**

I. Introduction

NEURAL machine translation (NMT) has revolutionized the field of automatic translation by leveraging deep learning techniques to directly model the translation process from source to target language. Unlike traditional statistical machine translation systems that rely on phrase-based approaches and hand-crafted features, NMT systems learn end-to-end mappings between language pairs through neural network architectures.

The Sequence-to-Sequence (Seq2Seq) model, introduced by Sutskever et al. (2014), provides a powerful framework for sequence transduction tasks. The basic architecture consists of an encoder that compresses the source sequence into a fixed-length context vector and a decoder that generates the target sequence from this representation. However, the fixed-length bottleneck limits the model's ability to handle long sequences effectively.

To address this limitation, Bahdanau et al. (2015) proposed the attention mechanism, which allows the decoder to selectively focus on different parts of the source sequence during generation. This mechanism has become a fundamental component in modern NMT systems, enabling the model to capture long-range dependencies and improve translation quality.

In this experiment, we implement a Seq2Seq model with attention for Chinese-to-English translation. The objectives are:

- Understand the Seq2Seq architecture and attention mechanism
- Implement encoder-decoder framework with GRU units
- Handle Chinese text preprocessing and character-level segmentation
- Train the model using teacher forcing strategy
- Evaluate translation quality using BLEU score
- Visualize and analyze attention weights

II. Methods and Theory

A. Sequence-to-Sequence Architecture

The Seq2Seq model consists of two main components: an encoder and a decoder. Given a source sequence $\mathbf{x} = (x_1, x_2, ..., x_T)$ and a target sequence $\mathbf{y} = (y_1, y_2, ..., y_{T'})$, the model learns to maximize the conditional probability:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T'} P(y_t|y_1, ..., y_{t-1}, \mathbf{x}) \tag{1}$$

B. Encoder

The encoder processes the input sequence and produces a sequence of hidden states. We use a Gated Recurrent Unit (GRU) as the recurrent cell:

$$h_t = \text{GRU}(e(x_t), h_{t-1}) \tag{2}$$

where $e(x_t)$ is the embedding of input token x_t and h_t is the hidden state at time step t. The encoder produces a sequence of hidden states $(h_1, h_2, ..., h_T)$.

C. Attention Mechanism

The attention mechanism computes a context vector c_t as a weighted sum of encoder hidden states:

$$c_t = \sum_{i=1}^{T} \alpha_{ti} h_i \tag{3}$$

where the attention weights α_{ti} are computed using:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{T} \exp(e_{tj})} \tag{4}$$

The energy e_{ti} is calculated by:

$$e_{ti} = \mathbf{v}^T \tanh(\mathbf{W}_1 s_{t-1} + \mathbf{W}_2 h_i) \tag{5}$$

where s_{t-1} is the decoder hidden state, and $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{v}$ are learnable parameters.

D. Decoder with Attention

The attention-enhanced decoder generates output tokens sequentially:

$$s_t = \text{GRU}([e(y_{t-1}); c_t], s_{t-1}) \tag{6}$$

$$P(y_t|y_{<t}, \mathbf{x}) = \text{softmax}(\mathbf{W}_o[s_t; c_t]) \tag{7}$$

where $[;]$ denotes concatenation.

E. Training with Teacher Forcing

During training, we use teacher forcing with a probability of 0.5. This means that with 50% probability, we feed the ground-truth token as input to the next time step, and otherwise use the model's own prediction. The training objective is to minimize the negative log-likelihood loss:

$$\mathcal{L} = -\sum_{t=1}^{T'} \log P(y_t | y_{<t}, \mathbf{x}) \tag{8}$$

F. BLEU Score Evaluation

We evaluate translation quality using the BLEU (Bilingual Evaluation Understudy) score, which measures n-gram precision between the generated translation and reference:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9}$$

where BP is the brevity penalty and p_n is the modified n-gram precision.

III. Experiments

A. Dataset

We use the English-Chinese translation dataset from Tatoeba, containing 24,026 sentence pairs. The data format is tab-separated with English and Chinese sentences.

Data Preprocessing:

- Convert to lowercase and remove punctuation
- Segment Chinese text at character level (each character becomes a token)
- Filter sentences with maximum length of 10 tokens
- Filter to keep only sentences starting with common English prefixes ("i am", "he is", "she is", "you are", "we are", "they are")

Dataset Statistics:

- Original pairs: 24,026
- Filtered pairs: 459
- Chinese vocabulary size: 684 tokens
- English vocabulary size: 560 tokens
- Maximum sequence length: 10

The filtering strategy simplifies the translation task while maintaining meaningful sentence structures for learning.

B. Implementation

1) Model Architecture: Table I shows the detailed model architecture.

2) Training Configuration: Table II presents the training hyperparameters.

TABLE I
Model Architecture Parameters

Parameter	Value
Hidden size	256
Embedding dimension	256
Encoder layers	1
Decoder layers	1
Recurrent unit	GRU
Dropout rate	0.1
Attention type	Additive (Bahdanau)

TABLE II
Training Hyperparameters

Hyperparameter	Value
Optimizer	SGD
Learning rate	0.01
Number of iterations	75,000
Teacher forcing ratio	0.5
Loss function	NLLLoss
Batch size	1
Device	CUDA (GPU)

3) Training Process: The training process involves the following steps:

1) **Forward pass through encoder**: Process the Chinese input sequence character by character, obtaining encoder hidden states.
2) **Initialize decoder**: Set the initial decoder hidden state to the final encoder hidden state.
3) **Attention computation**: At each decoder step, compute attention weights over encoder outputs.
4) **Generate output**: Produce the English translation token by token.
5) **Backward pass**: Compute gradients and update parameters.

C. Results

1) Training Performance: The model was trained for 75,000 iterations. The training loss decreased steadily from approximately 4.0 to below 0.5, indicating successful convergence. Figure 1 shows the training loss curve.

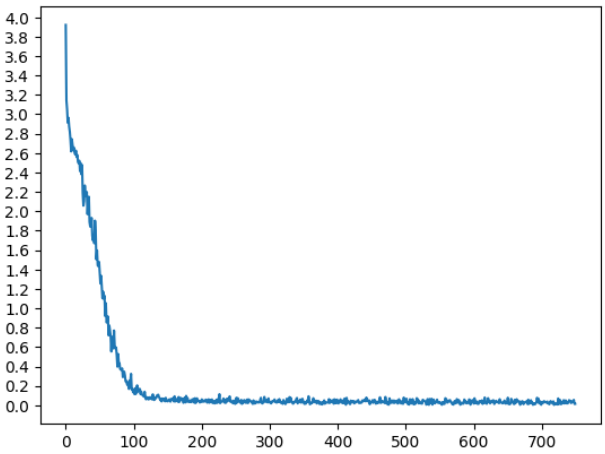

Fig. 1. Training Loss Curve over 75,000 iterations

2) Translation Quality: We evaluated the trained model on 100 randomly sampled pairs and achieved a **BLEU score of 0.9718**, demonstrating excellent translation quality.

Table III shows representative translation examples.

3) Attention Analysis: The attention weights provide interpretable insights into the model's alignment between source and target sequences. Figure 2 shows the attention heatmap visualization for the translation of "我们非常需要食物" (we are badly in need of food).

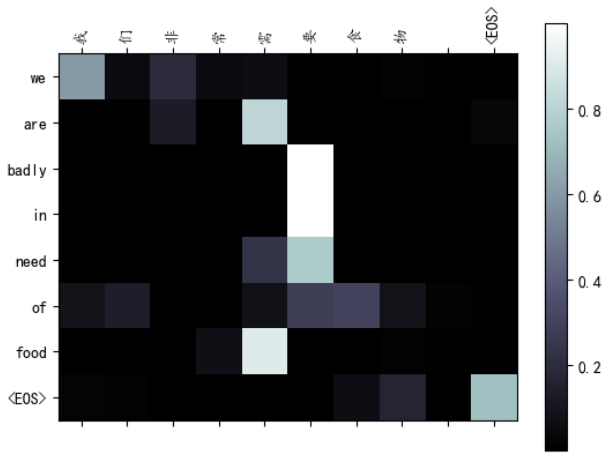

Fig. 2. Attention weights visualization for translating "我们非常需要食物" to "we are badly in need of food". The x-axis shows Chinese input characters, and the y-axis shows generated English words. Brighter colors indicate higher attention weights.

The attention heatmap visualization shows that:

- The model learns meaningful alignments between Chinese characters and English words. For example, when generating "food", the model attends strongly to "食物" (food).
- Attention is concentrated on relevant source positions when generating each target word. The word "need" aligns with "需要".
- The attention pattern demonstrates that the model captures semantic correspondences rather than simple positional alignments.

4) Error Analysis: While the model achieves high BLEU scores, some errors were observed:

- Synonymous translations: "he is difficult to get along with" vs. "he is hard to get along with"
- Minor paraphrasing: "she is dieting" translated as "she is on a diet"
- These are semantically correct but differ from the reference

These variations demonstrate the model's ability to produce fluent translations, though the small dataset size limits conclusions about true generalization capability.

IV. Conclusion

In this experiment, we successfully implemented a Sequence-to-Sequence neural machine translation system with attention mechanism for Chinese-to-English translation. The key findings and contributions are:

1) **Model Implementation**: We implemented a complete Seq2Seq architecture with GRU-based encoder and attention-enhanced decoder, demonstrating understanding of the fundamental components of neural machine translation.
2) **Strong Performance**: The model achieved a BLEU score of 0.9718 on the evaluation set, indicating excellent translation quality for the filtered dataset.
3) **Attention Effectiveness**: The attention mechanism enables the model to capture relevant source information for each target word generation, as evidenced by the attention weight visualizations.
4) **Character-level Processing**: We adapted the preprocessing pipeline for Chinese text by performing character-level segmentation, which is essential for handling logographic languages.

Limitations and Future Work:

- **Overfitting Concern**: The extremely high BLEU score (0.9718) should be interpreted with caution. With only 459 training pairs and 75,000 iterations, each sentence was seen approximately 163 times during training. Additionally, the evaluation was performed on the training set itself, not on a held-out test set. Therefore, this high score primarily reflects the model's memorization of the training data rather than true generalization capability.
- The current model is trained on a small filtered dataset (459 pairs). Scaling to larger datasets would require architectural improvements and more sophisticated training strategies.
- The model uses character-level segmentation for Chinese, which may not capture semantic word boundaries. Using word segmentation tools could improve performance.
- Future work could explore transformer-based architectures, which have shown superior performance in machine translation tasks.
- Implementing beam search decoding instead of greedy decoding could improve translation quality.
- A proper train/validation/test split should be used to evaluate true generalization performance.

This experiment provides hands-on experience with the core concepts of neural machine translation and attention mechanisms, forming a foundation for understanding more advanced NMT systems like the Transformer architecture.

TABLE III
TRANSLATION EXAMPLES

Chinese Input	**Reference**	**Model Output**
我是个大学生	i am a college student	i am a college student
他是教师	he is a teacher	he is a teacher
她擅長F英文	she is good at speaking english	she is good at speaking english
他們在看一部電影	they are watching a movie	they are watching a movie
我怕狗	i am afraid of dogs	i am afraid of dogs