LSTM-based Khmer Text Style Transfer Using Representation Learning

Pel Bunkhloem	e2021
Pheth Soriyuon	e2021
Phoeun Rajame	e2021
Phoeurn Kimhor	e2021
Phorn Sreypov	e2021
Yin Sambat	e2021

Institute of Technology of Cambodia

December 2025

1. Problem Statement

The task is to transfer text from a **source style** to a **target style** while preserving content. Formally, let:

- $X = (x_1, x_2, \dots, x_T)$ denote the input sequence in the source style.
- $Y = (y_1, y_2, \dots, y_{T'})$ denote the target sequence in the desired style.
- f_θ be the function implemented by our seq2seq LSTM, parameterized by θ.

We aim to model the conditional distribution:

$$P(Y \mid X; \theta) = \prod_{t=1}^{T'} P(y_t \mid y_1, \dots, y_{t-1}, X; \theta)$$

The objective is to find θ^* that maximizes the likelihood of the target sequences in the dataset $\mathcal{D}$:

$$\theta^* = \arg\max_\theta \sum_{(X,Y)\in\mathcal{D}} \log P(Y \mid X; \theta)$$

This is a **representation learning problem**, as the encoder LSTM learns hidden states h_t that represent the content of the input sequence in a way that can be decoded into the target style.

2. Proposed Solution

Model Architecture

We use an **encoder-decoder LSTM**:

Encoder LSTM

$$\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

- c_t is the **cell state** storing long-term information.
- h_t is the **hidden state** representing current content information.

Decoder LSTM

The decoder generates each token conditioned on previous outputs and the final encoder states:

$$P(y_t \mid y_{<t}, X) = \mathrm{Softmax}(W s_t + b)$$

where s_t is the decoder hidden state at step t.

Loss Function

The model is trained using **negative log-likelihood**:

$$\mathcal{L}(\theta) = -\sum_{(X,Y)\in\mathcal{D}} \sum_{t=1}^{T'} \log P(y_t \mid y_{<t}, X; \theta)$$

3. Challenges and Mitigations

1. **Long sequences:**
    - LSTM mitigates vanishing gradient issues, but very long sequences can still cause information loss.
    - *Mitigation:* Use bidirectional LSTM encoder to capture context from both directions.
2. **Content preservation vs style transfer:**
    - The decoder sometimes alters content when applying the target style.
    - *Mitigation:* Apply attention mechanism to let the decoder focus on relevant encoder hidden states.
3. **Limited parallel data:**
    - Perfect source-target pairs are scarce.
    - *Mitigation:* Pretrain the LSTM encoder-decoder on a language modeling task before fine-tuning on style transfer data.