Institute of Technology of Cambodia
Department of Applied Mathematics and Statistics

LSTM-based Khmer Text Style Transfer Using Representation Learning

Subject: Natural Language Processing (NLP)

Name of Students	ID
PEL Bunkhloem	e20201314
PHETH Soriyuon	e20210674
PHOEUN Rajame	e20211748
PHOEURN Kimhor	e20210823
PHORN Sreypov	e20210166
YIN Sambat	e20210138

Lecturers: **Dr. KHON Vanny (Course)**
Mr. TOUCH Sopheak (TP)

Academic Year 2025 - 2026

1 Introduction

1.1 Background and Motivation

The Khmer Language employs multiple sociolinguistic registers that reflect the social status and role of the person being addressed. Among these registers, រាជសព្ទ **(Royal language)** is of particular cultural and linguistic importance, as it is exclusively used when addressing or referring to the King. This register is characterized by specialized vocabulary and strict linguistic conventions that differ significantly from standard or plain Khmer usage.

In modern digital communication and language technology applications, proper use of royal language is often challenging. Most existing text processing systems treat Khmer as a single uniform style and do not account for sociolinguistic variation. As a result, incorrect or inappropriate register usage frequently occurs in written content, educational materials, and automated systems. This creates a strong motivation to develop computational approaches that can automatically convert plain Khmer text into its correct royal form while preserving the original meaning and cultural appropriateness.

1.2 Overview of Representation Learning

Representation learning is a fundamental concept in contemporary natural language processing (NLP), enabling models to learn meaningful linguistic features directly from data rather than relying on manually defined rules. Through distributed representations such as word and sentence embeddings, models can capture semantic relationships, contextual usage, and syntactic patterns in text.

Sequence-based deep learning models, particularly Recurrent Neural Networks (RNNs), are well suited for language transformation tasks because they process text sequentially and can model dependencies across words. In the context of stylistic transformation, representation learning allows the model to learn systematic differences between plain and royal Khmer, such as vocabulary substitutions and structural variations. By training on parallel corpora consisting of plain-text inputs and their royal-language equivalents, the model can learn to generate linguistically appropriate royal expressions while maintaining semantic consistency.

1.3 Scope and Objectives of the Study

The scope of this study is **strictly limited to the transformation of standard, plain Khmer text into the royal (រាជសព្ទ) register**. Other sociolinguistic registers, such as monk language (សង្ឃសព្ទ) or polite speech used for elders, are intentionally excluded to maintain a focused and well-defined research objective.

The specific objectives of this study are as follows:

- To develop an automated system that converts plain Khmer sentences into their corresponding royal-language forms.
- To apply deep learning–based sequence models for sociolinguistic style transfer.
- To ensure that the semantic meaning of the input text is preserved during transformation.
- To contribute to Khmer language processing research by addressing a low-resource language challenge and promoting culturally appropriate language use in computational systems.

This focused approach establishes a foundational framework that can be extended in future work to support additional Khmer sociolinguistic registers.

2 Problem Statement

2.1 Definition of the Problem

The Khmer language contains multiple sociolinguistic registers that vary according to the social status of the person being referenced. Among these registers, **Royal Khmer (រាជសព្ទ)** represents the highest and most formal level of language use. It differs substantially from normal Khmer text in terms of vocabulary selection, verb usage, and overall stylistic conventions.

The core problem addressed in this study is the development of an automated text transformation system capable of converting **normal Khmer text into Royal Khmer text**. This task can be viewed as a constrained form of linguistic style transfer, where the semantic meaning of the original sentence must be preserved while its linguistic style is transformed to conform to royal usage. To maintain a clear and manageable research focus, the scope of this work is **strictly limited to a single transformation direction: from normal text to royal text**. Other sociolinguistic transformations, such as monk language or elder-respectful language, are explicitly excluded from this study.

One of the primary challenges in addressing this problem is the lack of publicly available labeled datasets for Khmer sociolinguistic style transfer. As a result, all training data must be manually collected and annotated, which significantly constrains the size and diversity of the dataset. This low-resource setting presents additional difficulties for deep learning models, which typically require large amounts of data to achieve robust performance. Initial experimental results indicate that the model is not yet able to consistently generate correct royal-style output, often producing partially correct transformations or stylistically inaccurate expressions. These limitations highlight the complexity of modeling Royal Khmer under low-resource conditions.

2.2 Mathematical Formulation

The task is formulated as a **sequence-to-sequence style transfer problem**, where an input sentence in normal Khmer is transformed into its corresponding royal-style sentence while preserving semantic content.

Let:

- $X = (x_1, x_2, \dots, x_T)$ denote the input sequence in normal Khmer
- $Y = (y_1, y_2, \dots, y_{T'})$ denote the output sequence in Royal Khmer
- f_θ denote the encoder–decoder LSTM model parameterized by θ

The conditional probability of generating the target sequence given the input is modeled as:

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

The optimal parameters θ^* are obtained by maximizing the log-likelihood over the labeled dataset $\mathcal{D}$:

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

This formulation treats Khmer style transfer as a **representation learning problem**, where the encoder learns latent representations of the input sentence that can be decoded into the royal linguistic register.

2.3 Explanation of Terms and Notation

- X: Input sequence of tokens representing normal Khmer text
- Y: Target sequence of tokens representing Royal Khmer text
- x_t: The t-th token in the input sequence
- y_t: The t-th token in the output sequence
- T, T': Lengths of the input and output sequences

- $\mathcal{D}$: Manually labeled dataset of paired normal–royal Khmer sentences
- θ: Trainable parameters of the encoder–decoder LSTM model
- f_θ: Sequence-to-sequence function learned by the model
- h_t: Encoder hidden state representing contextual information at time step t

2.4 Remarks on Current Model Performance

Based on early experimental results, the model does not yet reliably perform correct style conversion. While some semantic content is preserved, royal-specific vocabulary and stylistic markers are frequently missing or incorrectly generated. These limitations are primarily attributed to the small size of the manually labeled dataset, limited training time, and the inherent complexity of Khmer royal language conventions.

3 Proposed Solution

We use an **encoder-decoder LSTM** for Khmer style transfer, adopting a **pretraining + fine-tuning approach** to improve performance on limited paired data.

3.1 Pretraining

Before fine-tuning, we **pretrained a single LSTM** on a large corpus of general Khmer text. The pretraining task was **self-reconstruction** ($X \rightarrow X$), meaning that given a sequence $X = (x_1, x_2, \dots, x_T)$, the model attempts to rewrite it.

- Pretraining loss:

$$\mathcal{L}_{pretrain}(\theta) = -\sum_{X \in \mathcal{C}} \sum_{t=1}^{T} \log P(x_t \mid x_1, \dots, x_{t-1}; \theta)$$

- Here, θ includes all LSTM weights and biases.
- Purpose: learn **robust Khmer token embeddings and sequence representations**.

The pretrained LSTM weights are then **used to initialize the encoder and/or decoder** of the style-transfer encoder-decoder model, giving the model a good starting point.

3.2 Model Architecture

3.2.1 Encoder LSTM

The encoder reads the input normal-style sentence X and generates **hidden states** h_t and **cell states** c_t:

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

- c_t stores long-term information.
- h_t represents the content of the input sentence.
- Encoder weights can be initialized with pretrained LSTM weights.

3.2.2 Decoder LSTM

The decoder generates the royal-style target sequence $Y = (y_1, \ldots, y_{T'})$ token by token:

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

- s_t = decoder hidden state at step t
- Each token depends on previous outputs and encoder representations.
- Decoder weights can also be initialized with pretrained LSTM weights.

3.3 Fine-Tuning for Style Transfer

The fine-tuning loss is **negative log-likelihood** over the paired dataset $\mathcal{D}$:

$$
\mathcal{L}_{finetune}(\theta) = -\sum_{(X,Y)\in\mathcal{D}} \sum_{t=1}^{T'} \log P(y_t \mid y_1, \ldots, y_{t-1}, X; \theta)
$$

- Fine-tuning adapts pretrained weights θ to the **style transfer task**.
- The model learns to **rewrite normal Khmer sentences into royal style**, while preserving content.