The Lao corpus and baseline models introduced in this study are publicly accessible at: https://github.com/HaHVU/HVULao_NLP

HVULao_NLP

HVULao_NLP is a project dedicated to sharing datasets and tools for Lao Natural Language Processing (NLP), developed and maintained by the research team at **Hung Vuong University (HVU), Phu Tho, Vietnam**.
This project is supported by **Hung Vuong University, Phu Tho, Vietnam**, with the aim of advancing research and applications in low-resource language processing, particularly for the Lao language.

📂 Datasets

This repository provides a semi-automatically constructed corpus consisting of Lao sentences that have been **word-segmented** and **part-of-speech (POS) tagged**. It is designed to support a wide range of NLP applications, including language modeling, sequence labeling, linguistic research, and the development of Lao language tools.

🔷 `Datatest1k/`

This folder contains the **test dataset**, consisting of 1,000 Lao sentences:

- `testorgin1000.txt`
  Original 1,000 Lao sentences in raw form (no segmentation or POS tags).
  Format: one sentence per line (UTF-8).

- `testsegsent_1000.txt`
  Word-segmented version aligned 1-to-1 with `testorgin1000.txt`.
  Format: each line is the segmented form of the same-index sentence; tokens separated by a single space.

- `testtag1k.json`
  → The same 1,000 sentences with **word segmentation** and **POS tagging**. These sentences are created using large language models (LLMs) like ChatGPT and then manually reviewed and corrected by native linguists.

🔷 `Datatrain10k/`

This folder contains the **training dataset**, consisting of 10,000 Lao sentences:

- `10ktrainorin.txt`
  Original 10,000 Lao sentences in raw form (no segmentation or POS tags).
  Format: one sentence per line (UTF-8).

- `10ksegmented.txt`
  Word-segmented version aligned 1-to-1 with `10ktrainorin.txt`.
  Format: each line is the segmented form of the same-index sentence; tokens separated by a single space.

- `10ktraintag.json`
  → The same 10,000 sentences with **word segmentation** and **POS tagging**, created using the same method as the test data.

> All data files are UTF-8 encoded and prepared for easy use in NLP pipelines.

📂 The Lao sentence segment tool:

This is a command-line tool for Lao word segmentation using a fine-tuned `transformers` model. It leverages Hugging Face's `transformers` library and PyTorch for efficient token classification.

Features

- Lao language word segmentation using a pre-trained and fine-tuned model
- Simple command-line usage
- GPU support for faster processing (if available)

📊 Evaluation Results

Model	Precision	Recall	F1-Score
LaoNLP	0.71	0.71	0.71
Flores	0.37	0.56	0.45
Our tool	**0.76**	**0.74**	**0.75**

Requirements

- Python 3.7+
- PyTorch
- Transformers (Hugging Face)
- A fine-tuned model (Please download at: https://huggingface.co/Tienha123/Segmenttool/tree/main)

You can install the dependencies with:

```
pip install torch transformers
```

Setup

1. Clone this repository (if from GitHub or Hugging Face).
2. Place your fine-tuned model folder in the same directory or update the `model_path` in the script accordingly.
3. Make sure your Python script is executable:

   ```
   chmod +x segment_lao.py
   ```

Input Format

- Input should be a **UTF-8 encoded text file**.
- Each line should contain a Lao sentence (without tokenization).

Usage

Run the script from the command line:

```
python3 segment_lao.py -i <input_file> -o <output_file>
```

Arguments

- `-i`, `--input`: Path to the input file (required)
- `-o`, `--output`: Path to the output file (required)

Example

```
python3 segment_lao.py -i ./data/lao_raw.txt -o ./output/lao_segmented.txt
```

How It Works

- Loads a fine-tuned token classification model (`lao_finetuned_10k`)
- Performs word segmentation based on B-WORD label prediction
- Outputs segmented sentences to the specified output file

Device Support

The tool automatically uses **GPU** if available. If not, it falls back to **CPU**.

Sample Output

Input:

```
ຂ້ອຍຮັກພາສາລາວ
```

Output:

```
ຂ້ອຍ ຮັກ ພາສາ ລາວ
```