Foundations of Natural Language Processing

Peking University, 2025

Assignment 4 Project 1:
Machine Translation for Low-Resource Languages

1. Directions

Please first read the **general instructions** of Assignment 4.

If you choose this project,

- Please submit your homework as a zip file through **Course**, which should include one report in PDF and your source code in Python. For LLM-based methods, you should additionally submit a JSON file containing the prompt for each testing instances
- Please include the score you achieved on the leaderboards of two sub-tasks in the report.
- The code should be paired with a README file describing dependencies, code structures, etc.

We will not simply grade your homework based on the model performance, but consider **the models you use**, the **novelty** of your method, the **workload**, and the **analysis** in your report.

If you graduate this summer, given that you have less time to complete the project, we will apply a more relaxed grading scale.

2. Task Description

In this project, you are going to build a machine translation system for a low-resource language, Zhuang (壮语). It is a Kra-Dai language spoken by Zhuang people (壮族) in the southern China. It is a challenging task because existing LLMs, even GPT and Gemini series, have limited abilities of this language.

Given limited data resources, you are required to translate the Zhuang sentences in the test set into Chinese. You are required to perform two sub-tasks:

Sub-Task 1: Simple and Controlled

Given a grammar book, you should translate simple Zhuang sentences/phrases into Chinese. For each sentence to be translated, it relies on one or several rules in the grammar book to be correctly translated. We additionally provide the necessary explanation for the words in the sentence, so that you do not need to consult to a dictionary.

Example:

Input:

A grammar book

A Zhuang sentence to be translated: *Gou dwg Vangz Gangh*.

Related words: *gou -> 我， Vangz Gangh -> 王刚*

Output:

The Chinese translation: *我是王刚。*

Sub-Task 2: Difficult and Realistic

Given a grammar book, a dictionary, and a few thousand parallel sentences, you should translate more complex Zhuang sentences. We do not provide the necessary explanation for the words in the sentence any more.

Example:

Input:

A grammar book, a dictionary, a few thousand parallel sentences

A Zhuang sentence to be translated: *Boux boux ma daengz lajmbwn couh miz cwyouz, cinhyenz caeuq genzli bouxboux bingzdaengj*.

Output:

The Chinese translation: *人人生而自由，在尊严和权利上一律平等。*

Method

There is **no constraint** on the method you use. You can implement your own model from scratch, finetune (large) language models, or use LLM APIs. If you use APIs, please use the qwen-max API as mentioned in the general instructions. Please clearly describe the method you use in the report.

The data for two sub-tasks are posted on two separate Kaggle competitions:

Sub-task 1: https://www.kaggle.com/t/5b442ae0a424454d862f43b5b8c5b384

Sub-Task 2: https://www.kaggle.com/t/cf84fc6677d94c9f83940a81da0da0f9

Evaluation

We use character-level BLEU scores with smoothing, implemented in the NLTK toolkit, to evaluate the translation quality.

The BLEU score is in the range of [0, 1].

```
>>> from nltk.translate.bleu_score import SmoothingFunction, sentence_bleu
>>> reference = ["你", "好", "吗", "?"]
>>> prediction = ["你", "好", "啊", "!"]
>>> smoothing = SmoothingFunction()
>>> sentence_bleu([reference], prediction, smoothing_function=smoothing.method1)
```

3. Resources

1. Here are some papers on low-resource translation, which might be useful for the project:

   Teaching Large Language Models an Unseen Language on the Fly

https://aclanthology.org/2024.findings-acl.519.pdf

A Benchmark for Learning to Translate a New Language from One Grammar Book https://openreview.net/pdf?id=tbVWug9f2h

Survey of Low-Resource Machine Translation https://aclanthology.org/2022.cl-3.6.pdf

Neural Machine Translation for the Indigenous Languages of the Americas: An Introduction https://aclanthology.org/2023.americasnlp-1.13.pdf

2. Here are some tutorials on prompting LLMs:

   https://www.promptingguide.ai/zh/introduction/basics

   https://learnprompting.org/zh-Hans/docs/basics/intro