ML2025
Bonus Competition

TA: Ulin Sanga 陳宥林
ntu-ml-2025-spring-ta@googlegroups.com
Deadline: 2025/06/06 23:59:59 (UTC+8)

Overview

- This is a translation task for the Puyuma language, which is a Formosan language spoken in Taiwan.
- The goal is to translate sentences from Mandarin Chinese to Puyuma and vice versa.

- ML 2025 Spring
- NTU COOL
- Team form
- Kaggle Competition
  - Join link
  - https://www.kaggle.com/t/bf7880c9aa95483580e0c7a500b122b2

Puyuma

- 卑南語
- Austronesian languages
- Native speaker: less than 4000 people
- Example sentences
  - ‘inava u ziya? 你好嗎？
  - i Ulin ku ngadan. 我的名字是 Ulin。

Task description

- Translate sentences from Puyuma to Mandarin.
  - 145 sentences
  - Example
    - “ai, marepana'uwa ta ziya.”
    - “alamu uturik mapiya, ikalalup murung za kiyakarukarunan.”
- Translate sentences from Mandarin to Puyuma.
  - 60 sentences
  - Example
    - “我沒有摩托車。”
    - “這包米多少錢？”

Metrics

- Syntactic level evaluation
- BLEU score
  - BLEU: a method for automatic evaluation of machine translation
- chrF
  - chrF: character n-gram F-score for automatic MT evaluation

Your final score is the average of these 4 score: (pyu: Puyuma, zh: Mandarin)

1. pyu to zh BLEU
2. zh to pyu BLEU
3. pyu to zh chrF
4. zh to pyu chrF

Dataset - Training set

- Download from Kaggle!
  - https://www.kaggle.com/competitions/ml-2025-bonus-competition/data
- CSV format
- Number of sentences
  - Puyuma, Mandarin parallel data: 1000
  - Puyuma, Mandarin, English data: 1000
  - Total: 2000 sentence pairs
- Example
  - “tratrima' ku za vurasi.”,“I want to buy sweet potatoes.”,“我要買地瓜。” (with English)
  - “tratrima' ku za vurasi.”,“我要買地瓜。” (without English)

Dataset - Lexicon

- Words and phrases in Puyuma, Mandarin (and English)
- **Not exhaustive**, but it can be a useful resource for building your translation system.
- Please note that **there might be duplicated entries** since the data are crawled from some online sources. You should take this into consideration when using the lexicon.
- Example
  - "sadeku","warm","暖和" (with English)
  - "Kinaveras","糯米飯" (without English)

Submission

- Submit your prediction file to Kaggle.
- For each ID in the test set, you must predict a translation.
  - The first 60 rows are Mandarin to Puyuma.
  - The rest are Puyuma to Mandarin.
  - Make sure you didn't put them in the wrong order.

```
ID,answer
1, 'inava u ziya?
2, ...
3, ...
...
61, 今天天氣真好。
62, ...
...
205, 我愛機器學習。
```

Submission

- Submit your technical report to NTU COOL.
- Only one member of the team have to upload the file.
- The file name must be **[your Kaggle Team name].pdf**
  - square bracket not included
- The content includes but not limited to the following
  - Models used
  - Methods
  - Proper citation

Participation

- This is a team project. You can have up to 5 members in your team.
- Please fill the following form about the member of your team.
  - https://forms.gle/x2DxuTug3iEivUyb8
- We will not deal with any of the workload distribution issue.
- No member changes are allowed. Your grades will be the same for the members in the same team. Find your teammates wisely.
- You have to fill the form even you are fighting on your own.
- Please note that the “Team Name” in the form must match your Kaggle Team Name.

Rules

- No manual translations allowed.
  - Do not search for the source of the dataset.
- Proprietary models are allowed.
  - Specify the model you used in technical report.

Grading

- No private leaderboard
- Total 10 points.
- Technical reports
  - Describe your method or system.
  - Will be graded based on clarity and novelty.
  - 3 points
- Baselines
  - Simple: 1 points
  - Medium: 2 points
  - Strong: 4 points
- Human evaluation
  - We will take the top 10 teams' submissions and conduct human evaluations.
  - Additional points will be given based on the preference of native speakers.
  - At most 3 points.

Regulations

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference.
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- Your final grade x 0.9 and get a score 0 for that homework if you violate any of the above rules first time (within a semester)
- Your will get F for the final grade if you violate any of the above rules multiple times (within a semester).
- Prof. Lee & TAs preserve the rights to change the rules & grades.

Some hints

- You may train a transformer-based machine translation model.
  - But the data is scarce, will the performance be good enough?
- You may utilize the power of LLMs.
  - Feed the grammar to LLMs?
  - Give the lexicons along with the sentence to be translated?
  - Agent may help.
  - RAG may help
- What LLMs might be good for the task?
  - NLLB?
  - GPT4.5?
  - Reasoning models?

Some resources

- Papers
  - Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?
  - A Benchmark for Learning to Translate a New Language from One Grammar Book
  - Teaching Large Language Models to Translate on Low-resource Languages with Textbook Prompting
- Puyuma related
  - 卑南語參考語法 A reference grammar of Puyuma, an Austronesian language of Taiwan
  - 族語辭典 https://glossary.ilrdf.org.tw/resources
  - 卑南語語法概論 https://alilin.cip.gov.tw/Book/417

Some insights about Puyuma language

- Agglutinative
  - terekuk (雞) -> puwa-terekuk (抓-雞) -> puwa-terekuk-an (雞舍)
- Keywords
  - Morpheme-rich
  - Morphological analyzer
  - Austronesian languages language model

Good Luck!