Applying_LLM_Demo

October 1, 2025

1 Applying LLM Technologies for Myanmar Language NLP

Purpose: This notebook demonstrates how LLMs and supporting algorithms can be applied to core Myanmar NLP tasks, while also highlighting the current limitations of LLMs for low-resource languages like Myanmar.

Written by Ye Kyaw Thu, LU Lab, Myanmar
Last updated: 30 Sept 2025
Demonstrated in a lecture at MyanmarSarYatWon, 1 Oct 2025

Email: ykt.nlp.ai@gmail.com

1.1 Technologies Used:

- Hugging Face Transformers (LLMs for generation, summarization, translation, NER).
- Sentence-Transformers (for semantic similarity).
- SymSpell (for spell checking).

ဒီ လက်တွေ့ သရုပ်ပြတဲ့နေရာမှာက အွန်လိုင်းမှာ အများသုံးလို့ရဖို့အတွက် ရဲထားကြတဲ့ မော်ဒယ်တွေကိုပဲ ယူသုံးပါမယ်။
ဒီမော်ဒယ်တွေက ကျွန်တော်တို့ Language Understanding Lab. က ဆောက်ထားတဲ့ မော်ဒယ်များ မဟုတ်ပါ။ ဒါပေမဲ့ ကျွန်တော်တို့ Lab က ရဲထားတဲ့ မြန်မာစာဒေတာတွေတော့ ပါကောင်း ပါနိုင်ပါတယ်။

1.2 –help

[1]:
```
!python mm_demo.py --help
```

```
2025-09-30 18:10:54.520333: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 18:10:54.528528: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759230654.537961  451701 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
```

```
E0000 00:00:1759230654.541097  451701 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759230654.548894  451701 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759230654.548906  451701 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759230654.548908  451701 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759230654.548909  451701 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 18:10:54.551112: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
usage: mm_demo.py [-h] [--device DEVICE]
                  {generate,summarize,translate,ner,sim,spellcheck} …

Myanmar NLP demo - generation / summarization / translation / ner / sim /
spellcheck

positional arguments:
  {generate,summarize,translate,ner,sim,spellcheck}
    generate            Text generation (Myanmar GPT or similar)
    summarize           Summarization task
    translate           Machine translation task
    ner                 Named entity recognition
    sim                 Sentence similarity (paraphrase check)
    spellcheck          Spell checking with SymSpell dictionary

options:
  -h, --help            show this help message and exit
  --device DEVICE       Device index for GPU (e.g. 0). Omit or -1 for CPU.
                        Defaults to GPU if available.
```

1.3 Text Generation

ဒီနေရာမှာ ဒီမို လုပ်ပြမှာက ရှိပြီးသား pretrained model တစ်ခုခုကိုသုံးပြီးတော့ text generation

(မြန်မာစာ စာကြောင်း တည်ဆောက်ခိုင်းတဲ့အလုပ်) ဘယ်လို လုပ်သလဲ။ ပေးလိုက်တဲ့ အစစာလုံးပေါ် မူတည်ပြီး ရလဒ်အနေနဲ့ကကော ဘယ်လိုရှိသလဲ ဆိုတာကိုပါ။

ပထမဆုံး အနေနဲ့ "မင်္ဂလာပါ" ဆိုတဲ့ ပုံမှန်သုံးနေတဲ့ မြန်မာစာကြောင်းတိုလေးကိုပဲ အစပြုပြီး GPT model က မြန်မာစာ စာကြောင်းကို ဘယ်လို အော်တိုမစ်တစ် ဆောက်ပေးသွားနိုင်သလဲ ဆိုတာကိုပါ။

[2]:
```
!time python mm_demo.py generate --text "မင်္ဂလာပါ" --max-length 80
```

```
2025-09-30 18:15:26.074048: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 18:15:26.082150: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759230926.091842  451819 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759230926.094970  451819 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759230926.102733  451819 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759230926.102744  451819 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759230926.102745  451819 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759230926.102747  451819 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 18:15:26.104967: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading generation model: jojo-ai-mst/MyanmarGPT (device=0)
tokenizer_config.json: 668kB [00:00, 119MB/s]
```

```
vocab.json: 999kB [00:00, 42.4MB/s]
merges.txt: 456kB [00:00, 42.1MB/s]
added_tokens.json: 83.2kB [00:00, 46.9MB/s]
special_tokens_map.json: 100%|          | 438/438 [00:00<00:00, 1.55MB/s]
config.json: 100%|                 | 925/925 [00:00<00:00, 3.38MB/s]
model.safetensors: 100%|            | 509M/509M [00:34<00:00, 14.6MB/s]
generation_config.json: 100%|           | 119/119 [00:00<00:00, 467kB/s]
Device set to use cuda:0
Truncation was not explicitly activated but `max_length` is provided a specific
value, please use `truncation=True` to explicitly truncate examples to max
length. Defaulting to 'longest_first' truncation strategy. If you encode pairs
of sequences (GLUE-style) with the tokenizer you can select this strategy more
precisely by providing a specific strategy to `truncation`.
The following generation flags are not valid and may be ignored: ['temperature',
'top_p']. Set `TRANSFORMERS_VERBOSITY=info` for more details.
Both `max_new_tokens` (=256) and `max_length`(=80) seem to have been set.
`max_new_tokens` will take precedence. Please refer to the documentation for
more information.
(https://huggingface.co/docs/transformers/main/en/main_classes/text_generation)

--- GENERATION ---

--- sequence 1 ---
မင်္ဂလာပါတီနှင့် ကာလကြာမြင့်စွာ တိုက်ပွဲဝင်ခဲ့ရသည်။ ၁၉၄၅ ခုနှစ် ဒုတိယ
ကမ္ဘာစစ်ကြီး ပြီးနောက် ချန်ကေရှိတ်သည် တရုတ်ကွန်မြူနစ်များကို ချေမှုန်းနိုင်ရန်
ကြိုးပမ်းခဲ့သော်လည်း ကွန်မြူနစ်ဘက်မှ ဆိုဗီယက် ရုရှားတို့က ထောက်ခံအားပေးခဲ့ရာ
ကွန်မြူနစ်ကို စစ်ရှုံးခဲ့သည်။ ၁၉၄၉ ခုနှစ်တွင် ချန်ကေရှိတ်သည် ထိုင်ဝမ်သို့
တိမ်းရှောင်ခဲ့ပြီး ကွန်မြူနစ်လက်ထဲမှ တရုတ်ပြည်မကြီးအား ပြန်လည်တိုက်ယူနိုင်ရန်
ထိုင်ဝမ်တွင် ခြေကုတ်ယူခဲ့သည်။ သို့သော်လည်း သူကွယ်လွန်သည်အထိ အကောင်အထည်
မဖော်နိုင်ခဲ့ချေ။ ချန်ကေရှိတ်သည် ထိုင်ဝမ်သို့ ရောက်ရှိပြီးနောက်
တရုတ်သမ္မတနိုင်ငံ၏ သမ္မတအဖြစ် အုပ်စိုးခဲ့သည်မှာ သူကွယ်လွန်သည့် ၁၉၇၅ ခုနှစ်အထိ
ပင်ဖြစ်သည်။




real    0m44.631s
user    0m12.180s
sys     0m2.372s
```

အထက်မှာ မြင်ရတဲ့အတိုင်း စာကြောင်းကိုတော့ ဆောက်ပေးနိုင်ပါတယ်။ ဒါပေမဲ့ ဖတ်ကြည့်ရင် အဓိပ္ပါယ် အရရော စာလုံး၊ စာကြောင်း ဆက်စပ်မှုအရရော လိုအပ်ချက်တွေ ရှိနေသေးတာကို မြင်ကြပါလိမ့်မယ်။ အထက်ပါ စာတွေကို ထုတ်ဖို့ စုစုပေါင်း ၄၄ စက္ကန့် ကြာပါတယ်။

စာကြောင်း အရှည်ကိုတော့ –max-length နဲ့ ထိန်းလို့ ရပါတယ်။
ဒီတစ်ခေါက် –max-length ကို ပိုရှည်ရှည်ထားပြီး ထုတ်ခိုင်းကြည့်ပါမယ်။ အစစာလုံးကိုလည်း “ပုဂံပြည်”

ဆိုတာနဲ့ စာကြည့်ပါမယ်။

[3]:
```
!time python mm_demo.py --device 0 generate --text "ပုဂံပြည်" --max-length 200
```

```
2025-09-30 18:21:54.334416: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 18:21:54.343048: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759231314.352620  452530 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759231314.355765  452530 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759231314.363746  452530 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759231314.363757  452530 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759231314.363759  452530 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759231314.363760  452530 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 18:21:54.366014: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading generation model: jojo-ai-mst/MyanmarGPT (device=0)
Device set to use cuda:0
Truncation was not explicitly activated but `max_length` is provided a specific
value, please use `truncation=True` to explicitly truncate examples to max
length. Defaulting to 'longest_first' truncation strategy. If you encode pairs
of sequences (GLUE-style) with the tokenizer you can select this strategy more
```

```
precisely by providing a specific strategy to `truncation`.
The following generation flags are not valid and may be ignored: ['temperature',
'top_p']. Set `TRANSFORMERS_VERBOSITY=info` for more details.
Both `max_new_tokens` (=256) and `max_length`(=200) seem to have been set.
`max_new_tokens` will take precedence. Please refer to the documentation for
more information.
(https://huggingface.co/docs/transformers/main/en/main_classes/text_generation)

--- GENERATION ---

--- sequence 1 ---
ပုဂံပြည်သည် ၎င်း၏ ဥရောပတိုက်၊ တောင်ဘက်တွင် မြေထဲပင်လယ်ရှိပြီး အရှေ့တောင်ဘက်တွင်
ပင်လယ်နက်နှင့် ၎င်းနှင့်ဆက်သွယ်ထားသော ရေလမ်းကြောင်း ရှိသည်။
ဥရောပသည် ကမ္ဘာပေါ်ရှိတိုက်ကြီးများတွင် မြေမျက်နှာပြင်အားဖြင့် ဒုတိယမြောက်
အသေးဆုံးဖြစ်ပြီး ၁၀, ၁၈၀, ၀၀၀ စတုရန်းကီလိုမီတာ (၃, ၉၃၀, ၀၀၀ စတုရန်းမိုင်) သို့မဟုတ်
ကမ္ဘာ့မျက်နှာပြင်၏ ၂ရာခိုင်နှုန်း နှင့် ကမ္ဘာ့မြေမျက်နှာပြင်၏ ၆. ၈
ရာခိုင်နှုန်းကို နေရာယူထားသည်။ ဥရောပတွင် နိုင်ငံပေါင်း ၅၀ ခန့်ရှိပြီး
ရုရှားနိုင်ငံသည် လူဦးရေအားဖြင့်သော်လည်းကောင်း၊ အကျယ်အဝန်းအားဖြင့်သော်လည်းကောင်း
အကြီးဆုံးဖြစ်ပြီး အသေးဆုံးမှာ ဗာတီကန်စီးတီး ဖြစ်သည်။ ဥရောပသည် အာရှ နှင့်
အာဖရိကပြီးလျှင် တတိယမြောက် လူဦးရေ အများဆုံးတိုက် ဖြစ်သည်။


real    0m7.034s
user    0m10.634s
sys     0m0.751s
```

GPT မော်ဒယ် သို့မဟုတ် Large Language Model (LLM) က ရှိုးနေတယ် ဆိုတာကိုတော့ ကောင်းကောင်း မြင်ကြရပါလိမ့်မယ်။ :)
ဒီနေရာမှာ အသုံးပြုခဲ့တဲ့ မော်ဒယ်က MyanmarGPT ဆိုတဲ့ မော်ဒယ်ပါ။
HuggingFace Link: https://huggingface.co/jojo-ai-mst/MyanmarGPT

တစ်ခုသိစေချင်တာက လက်ရှိ ခင်ဗျားတို့ အသုံးပြုနေကြတဲ့ ChatGPT, Gemini, DeepSeek မော်ဒယ်တွေက တကယ်က GPT မော်ဒယ် အကြီးစားတွေပါပဲ။ ကျွန်တော်တို့သာ မြန်မာစာ ဒေတာတွေကို အများကြီး အချိန်ယူပြီး ပြင်ဆင်သွားမယ်ဆိုရင် လက်ရှိ ရှိနေတဲ့ မော်ဒယ်တွေထက် အများကြီးပိုကောင်းတဲ့ မော်ဒယ်တွေကို ဆောက်နိုင်ကြပါလိမ့်မယ်။

GPT မော်ဒယ်ကို စိတ်ဝင်စားတဲ့ သူတွေအနေနဲ့က ကျွန်တော်တို့ Lab က ကဗျာ ကောပတ်စ်ကို အတိုင်းအတာတစ်ခုအထိ ပြင်ဆင်ပြီး experiment လုပ်ပြထားတာကိုလည်း အောက်ပါလင့် ကနေ လေ့လာနိုင်ပါတယ်။

myPoetry: https://github.com/ye-kyaw-thu/myPoetry

1.4 Summarization

ဒီတခါတော့ အင်္ဂလိပ်လို text summarization လို့ခေါ်တဲ့ သတင်းဆောင်းပါး၊ ဝတ္ထုတို့ကို အတိုချုပ်ခိုင်းတဲ့ NLP အလုပ်ကို မြန်မာစာအတွက် လုပ်ခိုင်းကြည့်ပါမယ်။ mT5_multilingual_XLSum ဆိုတဲ့ မော်ဒယ်ကိုသုံးကြည့်ပါမယ်။ Multilingual မော်ဒယ်ဖြစ်ပြီး မြန်မာစာကိုလည်း အတိုင်းအတာတခုထိ

နားလည်အောင် training လုပ်ထားပါတယ်။
HaggingFace Link: https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum

မြန်မာသတင်းက BBC Burmese ကနေကော်ပီကူးယူထားတာပါ။
News Link: https://www.bbc.com/burmese/live/cj4y0l0vv1xt

သတင်းခေါင်းစဉ်ကို တမင်တကာ ဖြုတ်ထားခဲ့ပါတယ်။
Summarization လုပ်ခိုင်းမယ့် သတင်းအပြည့်အစုံက အောက်ပါအတိုင်းပါ။

1.4.1 Filename: article-1.txt

မန္တလေး၊ ပြင်ဦးလွင်မြို့၊ အုန်းချောရွာအနီး ၈ ထပ် ရေတံခွန်မှာ တောင်ကျရေနဲ့ မျောပါခဲ့သူတွေထဲက ပျောက်ဆုံးနေသူ ၂ ဦး ဒီကနေ့ စက်တင်ဘာ ၃၀ ရက်မှာ ရှာဖွေကယ်ဆယ်ရေး ဆက်လုပ်မယ်လို့မြန်မာနိုင်ငံလူမှုကယ်ဆယ်ရေးအဖွဲ့ (မန္တလေးတိုင်းရုံး)က တာဝန်ရှိသူတစ်ဦးက ဘီဘီစီကို ပြောပါတယ်။

စက်တင်ဘာ ၂၉ ရက် ညနေပိုင်းက တောင်ကျချောင်းရေကြောင့် ၈ ထပ် ရေတံခွန်မှာ ရေကစားနေသူ ၅ ဦး ရေစီးနဲ့မျောပါပျောက်ဆုံးခဲ့တဲ့အဲဒီဖြစ်စဉ်မှာ အသက် ၂၀ ကျော်အရွယ် မြန်မာ အမျိုးသမီး ၂ ဦး သေဆုံးပြီး ရှာရှားနိုင်ငံသား ၁ ဦးကို ကယ်ဆယ်နိုင်ခဲ့တယ်လို့ မြန်မာနိုင်ငံလူမှုကယ်ဆယ်ရေးအဖွဲ့ (မန္တလေးတိုင်းရုံး)ရဲ့ အချက်အလက်အရ သိရှိရပါတယ်။

ကယ်ဆယ်ခဲ့တဲ့ရှာရှားနိုင်ငံသားက ညာခြေကျိုး ထိခိုက်ဒဏ်ရာရရှိထားပြီး အသက် ၄၃ နှစ်အရွယ်လို့ ဆိုပါတယ်။

“ရှာရှားနိုင်ငံသားက ညာခြေထောက်ပဲကျိုးပါတယ်။ အသက်အန္တရာယ် မစိုးရိမ်ရဘူး။ ကျန်တဲ့ အသက် ၂၀ ကျော်နဲ့ ၂၅ နှစ်အရွယ် အမျိုးသမီး ၂ ဦးကတော့ ဆုံးသွားပါတယ်။ နောက်ထပ် ၂ ဦးကို ဒီမနက်ဆက်ရှာပါမယ်။ သူတို့က ၁၅ ဦးအဖွဲ့လို့သိရတယ်။ ရေထဲဆင်းတဲ့ ၅ ဦးပဲ မျောပါသွားတာပါ။ကျန်သူတွေက ဘေးကင်းသွားတယ်” လို့ မန္တလေးလူမှုကယ်ဆယ်ရေးရုံး က တာဝန်ရှိသူကရှင်းပြပါတယ်။

ပျောက်ဆုံး နေသူ ၂ ဦးဟာမြန်မာ အမျိုးသမီး ၂ ဦးလို့ ကနဦးသိထားရပါတယ်။

1.4.2 Summarization-1

[4]:
```
!time python mm_demo.py --device 0 summarize --text-file ./data/article-1.txt␣
↪--max-length 10
```

```
2025-09-30 18:45:41.344634: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 18:45:41.352848: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759232741.362258  453348 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759232741.365390  453348 cuda_blas.cc:1407] Unable to register
```

```
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759232741.373118  453348 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759232741.373129  453348 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759232741.373130  453348 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759232741.373132  453348 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 18:45:41.375337: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading summarization model: csebuetnlp/mT5_multilingual_XLSum (device=0)
You are using the default legacy behaviour of the <class
'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expected, and
simply means that the `legacy` (previous) behavior will be used so nothing
changes for you. If you want to use the new behaviour, set `legacy=False`. This
should only be set if you understand what it means, and thoroughly read the
reason why this was added as explained in
https://github.com/huggingface/transformers/pull/24565
/home/ye/.local/lib/python3.12/site-
packages/transformers/convert_slow_tokenizer.py:564: UserWarning: The
sentencepiece tokenizer that you are converting to a fast tokenizer uses the
byte fallback option which is not implemented in the fast tokenizers. In
practice this means that the fast version of the tokenizer can produce unknown
tokens whereas the sentencepiece version would have converted these unknown
tokens into a sequence of byte tokens matching the original piece of text.
  warnings.warn(
Device set to use cuda:0

--- SUMMARY ---

မန္တလေး၊ ပြင်ဦးလွင်မြို့၊

real    0m10.497s
```

```
user    0m12.354s
sys     0m2.959s
```

–max-length ကို 50 လို့ တိုးပေးကြည့်ရအောင်။

[5]:
```
! time python mm_demo.py summarize --text-file ./data/article-1.txt␣
↪--max-length 50
```

```
2025-09-30 18:46:58.316780: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 18:46:58.325134: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759232818.334843  453456 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759232818.338068  453456 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759232818.345991  453456 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759232818.346002  453456 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759232818.346004  453456 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759232818.346005  453456 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 18:46:58.348339: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading summarization model: csebuetnlp/mT5_multilingual_XLSum (device=0)
You are using the default legacy behaviour of the <class
```

```
'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expected, and
simply means that the `legacy` (previous) behavior will be used so nothing
changes for you. If you want to use the new behaviour, set `legacy=False`. This
should only be set if you understand what it means, and thoroughly read the
reason why this was added as explained in
https://github.com/huggingface/transformers/pull/24565
/home/ye/.local/lib/python3.12/site-
packages/transformers/convert_slow_tokenizer.py:564: UserWarning: The
sentencepiece tokenizer that you are converting to a fast tokenizer uses the
byte fallback option which is not implemented in the fast tokenizers. In
practice this means that the fast version of the tokenizer can produce unknown
tokens whereas the sentencepiece version would have converted these unknown
tokens into a sequence of byte tokens matching the original piece of text.
  warnings.warn(
Device set to use cuda:0

--- SUMMARY ---
```

မန္တလေး၊ ပြင်ဦးလွင်မြို့၊ အုန်းချောရွာအနီး ၈ ထပ် ရေတံခွန်မှာ တောင်ကျရေနဲ့ မျောပါခဲ့သူတွေထဲက ပျောက်ဆုံးနေသူ ၂ ဦး သေဆုံးပြီး ရ

```
real    0m10.275s
user    0m12.351s
sys     0m2.636s
```

တကယ်က အတိုချုပ်တဲ့ အလုပ်ကို ဒီမော်ဒယ်က မြန်မာစာအတွက် အလုပ်မလုပ်ဘူးလို့ ကျွန်တော်ကတော့ နားလည်ပါတယ်။ ဘာကြောင့်လည်း ဆိုတော့ ဒီ မော်ဒယ်က လူလည်ကျပြီး ဖတ်ခိုင်းထားတဲ့ input text ရဲ့ ထိပ်ဆုံးစာကြောင်းကိုပဲ ပြန်ခန့်မှန်းနေတာမို့ပါ။ ဟုတ်ပါတယ် သတင်းခေါင်းစဉ်တို့ ဆောင်းပါးနာမည်တို့က စာတွေရဲ့ ထိပ်ဆုံးမှာပဲ ရှိကြတာကိုး။ ကျွန်တော်တို့ input ပေးလိုက်တဲ့ ဖိုင်ရဲ့ ထိပ်ဆုံး စာကြောင်းက အောက်ပါအတိုင်းပါ။ မော်ဒယ်က ထွက်လာတာနဲ့ နှိုင်းယှဉ်ကြည့်လို့ ရအောင် ပြထားတာပါ။

မန္တလေး၊ ပြင်ဦးလွင်မြို့၊ အုန်းချောရွာအနီး ၈ ထပ် ရေတံခွန်မှာ တောင်ကျရေနဲ့ မျောပါခဲ့သူတွေထဲက ပျောက်ဆုံးနေသူ ၂ ဦး ဒီကနေ့ စက်တင်ဘာ ၃၀ ရက်မှာ ရှာဖွေကယ်ဆယ်ရေး ဆက်လုပ်မယ်လို့မြန်မာနိုင်ငံလူမှုကယ်ဆယ်ရေးအဖွဲ့ (မန္တလေးတိုင်းရုံး)က တာဝန်ရှိသူတစ်ဦးက ဘီဘီစီကို ပြောပါတယ်။

1.4.3 Summarization-2

ဒီတခါ input လုပ်ပေးမယ့် သတင်းတိုကတော့ အောက်ပါအတိုင်းပါ။

ထိုင်းနိုင်ငံ၊ ချင်းမိုင်မြို့က အိမ်မွေးတိရစ္ဆာန်ချစ်သူတွေ စုပေါင်းကျင်းပခဲ့တဲ့ “ခွေးလေးများနှင့် မိတ်ဆွေများ ဥယျာဉ်ပါတီ” ကို စက်တင်ဘာ ၁၉ ရက်က Jing Jai Market မှာ ပြုလုပ်ခဲ့ပါတယ်။

ဒီပွဲကို စက်တင်ဘာ ၁၈ ရက်ကစပြီး ၁၉ ရက်ထိ ၂ ရက်ကြာ ကျင်းပခဲ့တာလည်း ဖြစ်ပါတယ်။

ဒီပွဲမှာ အိမ်မွေးတိရစ္ဆာန်ပိုင်ရှင်တွေနဲ့ သူတို့ရဲ့ အိမ်မွေးတိရစ္ဆာန် အဖော်လေးတွေအတွက် ပျော်ရွှင်စရာကောင်းတဲ့ ပွဲတော်တခုအဖြစ် ဖန်တီးပေးခဲ့ပြီး အိမ်မွေးတိရစ္ဆာန် ကျန်းမာရေးနဲ့ အိမ်မွေးတိရစ္ဆာန်လေးတွေနဲ့ ပိုင်ရှင်တွေရဲ့ ပျော်ရွှင်စရာရပ်ဝန်းတခု တည်ဆောက်ပေးဖို့ ရည်ရွယ်တာလည်း ဖြစ်ပါတယ်။

ဒီပွဲကို တက်ရောက်လာသူတွေဟာ အခမဲ့ အိမ်မွေးတိရစ္ဆာန် ကျန်းမာရေးစစ်ဆေးတာနဲ့ ဖန်တီးမှုဆိုင်ရာ အလုပ်ရုံဆွေးနွေးပွဲ အပါအဝင် လှုပ်ရှားမှု အစုံအလင်ကို ပျော်ရွှင်စွာ ပါဝင်ဆင်နွှဲခဲ့ကြပါတယ်။

ဒီပွဲကို တက်ရောက်ခဲ့တဲ့ အိမ်မွေးတိရစ္ဆာန်ပိုင်ရှင်တယောက်ဖြစ်သူ ဂျင်(မ်) က သူ့ရဲ့အတွေ့အကြုံကို မျှဝေခဲ့ပါတယ်။

“ထိုင်းနိုင်ငံ၊ အထူးသဖြင့် ချင်းမိုင်မှာ အိမ်မွေးတိရစ္ဆာန်ပွဲလေးတွေ၊ ဥယျာဉ်ထဲမှာလုပ်တဲ့ ပါတီလေးတွေ ဒီလိုမျိုးပွဲတွေက အနည်းငယ်ပဲ ရှိသေးတာဆိုတော့ ဒါဟာ အခွင့်အရေးကောင်းတခုလို့ ကျနော် ထင်ပါတယ်။ ဒါက တခြားအိမ်မွေးတိရစ္ဆာန် မွေးထားတဲ့သူတွေကို ပိုပြီးသိရှိစေဖို့နဲ့ ဒီလိုပွဲမျိုးဟာ ဒီနေ့ခေတ် လူအဖွဲ့အစည်းအတွက် ကောင်းမွန်တဲ့ အသိုင်းအဝိုင်းတခု ဖြစ်လာနိုင်တယ် ဆိုတာကို သိရှိစေဖို့ အခွင့်အရေးကောင်းတခု ဖြစ်စေပါတယ်။”

ဒီပွဲမှာ အိမ်မွေးတိရစ္ဆာန်သုံးပစ္စည်းမျိုးစုံနဲ့ သာမန်ဆိုင်တွေမှာ ရှာရခဲတဲ့ ဇီဝခွေးစာတွေကိုပါ ရောင်းချပေးတဲ့ ဆိုင်ခန်းတွေလည်း ပါဝင်ခဲ့ပါတယ်။

[6]:
```
! time python mm_demo.py summarize --text-file ./data/article-2.txt␣
↪--max-length 50
```

```
2025-09-30 18:57:24.450842: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 18:57:24.459015: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759233444.468453  453681 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759233444.471631  453681 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759233444.479586  453681 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759233444.479596  453681 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759233444.479598  453681 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759233444.479600  453681 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 18:57:24.481862: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
```

```
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading summarization model: csebuetnlp/mT5_multilingual_XLSum (device=0)
You are using the default legacy behaviour of the <class
'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expected, and
simply means that the `legacy` (previous) behavior will be used so nothing
changes for you. If you want to use the new behaviour, set `legacy=False`. This
should only be set if you understand what it means, and thoroughly read the
reason why this was added as explained in
https://github.com/huggingface/transformers/pull/24565
/home/ye/.local/lib/python3.12/site-
packages/transformers/convert_slow_tokenizer.py:564: UserWarning: The
sentencepiece tokenizer that you are converting to a fast tokenizer uses the
byte fallback option which is not implemented in the fast tokenizers. In
practice this means that the fast version of the tokenizer can produce unknown
tokens whereas the sentencepiece version would have converted these unknown
tokens into a sequence of byte tokens matching the original piece of text.
  warnings.warn(
Device set to use cuda:0

--- SUMMARY ---
```

ထိုင်းနိုင်ငံမှာ အိမ်မွေးတိရစ္ဆာန်ချစ်သူတွေ စုပေါင်းကျင်းပခဲ့တဲ့ "ခွေးလေးတွေနဲ့ မိတ်ဆွေများ ဥယျာဉ်ပါတီ"

```
real    0m14.429s
user    0m12.380s
sys     0m2.513s
```

ဒီတခါတော့ မော်ဒယ်က အလုပ်လုပ်ပေးတယ်လို့ ကျွန်တော်တို့ ယူဆလို့ ရနိုင်ပါလိမ့်မယ်။ :)
တကယ်တမ်း အလုပ်လုပ်တယ် မလုပ်ဘူး ဆိုတာကို အကဲဖြတ်ဖို့က အလွယ်ကြီး မဟုတ်ပါဘူး။ ဒါပေမဲ့ NLP သမားတွေမှာတော့ သင့်လျော်တဲ့ အကဲဖြတ်နည်းလမ်းတွေကို သုံးပြီး အသေးစိတ် အကဲဖြတ်ကြရပါတယ်။ ဒီနေရာမှာတော့ text summarization သရုပ်ပြတာကို ဒီလောက်နဲ့ပဲ ထားလိုက်ကြရအောင်။

News link: https://burmese.dvb.no/post/725349

မြန်မာစာနဲ့ ပတ်သက်တဲ့ Text Summarization သုတေသန အလုပ်တချို့ကို ကျွန်တော့်အနေနဲ့ လက်ရှိမှာ ထိုင်းတက္ကသိုလ် တစ်ခုဖြစ်တဲ့ SIIT (Sirindhorn International Institute of Technology) မှာ ပါရဂူဘွဲ့ကျောင်းသူ လှိုင်မြတ်နွယ် နဲ့အတူ လုပ်ဆောင်လျက်ရှိပါတယ်။

1.5 Translation

ကျွန်တော်တို့ရဲ့ Natural Language Processing မှာတော့ စက်ကို ဘာသာစကားတစ်ခုကနေ တခြားမတူတဲ့

ဘာသာစကားတစ်ခုကို ဘာသာပြန်ခိုင်းတဲ့ အလုပ်ကလည်း အရေးကြီးတဲ့ သုတေသန အလုပ်တစ်ခုပါ။ ဘာကြောင့်လဲ ဆိုတော့ ဘာသာပြန်တဲ့ အလုပ်က လူတွေအတွက်တောင် အင်မတန်ခက်ခဲတဲ့ အလုပ်ဖြစ်လို့ပါ။ အင်တာဗျူးတို့ တရားရုံးတို့ အစည်းအဝေးတို့မှာ တိုက်ရိုက် ဘာသာပြန်ပေးဖို့ဆိုရင် လူတစ်ယောက်တည်းက အချိန်အကြာကြီး မလုပ်နိုင်ပါဘူး။ ခေါင်းတအား ပင်ပန်းလို့ပါ။ အထူးသဖြင့် အဲဒီလို live translation အတွက်ဆိုရင် R&D အနေနဲ့က လုပ်စရာတွေ အများကြီး ကျန်ပါသေးတယ်။

ဒီတခါမှာလည်း ကျွန်တော်တို့ မြန်မာလူမျိုးအားလုံးလိုလို သိကြတဲ့ Facebook (သို့မဟုတ်) Meta ကုမ္ပဏီရဲ့ NLP/AI သုတေသနပညာရှင်တွေက ဆောက်ထားတဲ့ pretrained translation model တခုကိုသုံးပြီး တခြားဘာသာကနေ မြန်မာကို ပြန်တာ၊ မြန်မာကနေ တခြားဘာသာ တစ်ခုကို ပြန်တာတွေကို လုပ်ပြပါမယ်။

HuggingFace Link: https://huggingface.co/facebook/nllb-200-distilled-600M

အရင်ဆုံး အင်္ဂလိပ် ကနေ ဗမာလို နှုတ်ဆက်တဲ့ စာကြောင်းအတိုလေးကိုပဲ ဘာသာပြန်ကြည့်ကြရအောင်။

[7]:
```
!time python mm_demo.py --device 0 translate --src eng_Latn --tgt mya_Mymr␣
↪--text "Hello, how are you?"
```

```
2025-09-30 19:14:31.558052: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 19:14:31.566464: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759234471.576247  454094 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759234471.579457  454094 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759234471.587366  454094 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759234471.587376  454094 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759234471.587378  454094 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759234471.587379  454094 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 19:14:31.589579: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
```

```
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading translation model: facebook/nllb-200-distilled-600M (device=0)
src=eng_Latn tgt=mya_Mymr
Device set to use cuda:0

--- TRANSLATION ---

မင်္ဂလာပါ။  မင်းဘယ်လိုလဲ။

real    0m8.571s
user    0m10.314s
sys     0m2.157s
```

“မင်းဘယ်လိုလဲ” ဆိုတာထက် “နေကောင်းလား”၊ “စားပြီးပြီလား” ဆိုတာမျိုး ဖြစ်သင့်ပါတယ်။

ဒီတခါတော့ I am a cook. ဆိုတဲ့ စာကြောင်းကို ဗမာလို ဘာသာပြန်ခိုင်းကြည့်ရအောင်။

[9]:
```
!time python mm_demo.py --device 0 translate --src eng_Latn --tgt mya_Mymr␣
↪--text "I am a cook."
```

```
2025-09-30 19:22:05.036246: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 19:22:05.044693: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759234925.054698  454349 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759234925.058067  454349 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759234925.066233  454349 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759234925.066246  454349 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759234925.066248  454349 computation_placer.cc:177] computation
```

```
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759234925.066249  454349 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 19:22:05.068660: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading translation model: facebook/nllb-200-distilled-600M (device=0)
src=eng_Latn tgt=mya_Mymr
Device set to use cuda:0

--- TRANSLATION ---

ကျွန်မက ချက်ပြုတ်သူပါ။

real    0m8.593s
user    0m10.388s
sys     0m2.175s
```

အဓိပ္ပါယ်တော့ မိပါတယ်။ ဒါပေမဲ့ “ထမင်းချက်”၊ “စားဖိုမှူး” လို့ ဘာသာပြန်ပေးနိုင်ရင် ပိုကောင်းပါတယ်။

နောက်ထပ် pretrained model ကိုသုံးတဲ့ machine translation ဥပမာအနေနဲ့ ဗမာစာ ကနေ အင်္ဂလိပ်လို ပြန်ခိုင်းကြည့်မယ်။
“မန္တလေးမှာ ပြောပြောနေကြတဲ့ တစ်ပြ ဆိုတာ ဘာလဲဗျ” ဆိုတဲ့ စာကြောင်းဆိုရင် စက်အနေနဲ့က ဘာသာပြန်ပေးနိုင်ဖို့ ခက်ပါလိမ့်မယ်။ လက်တွေ့ စမ်းကြည့်ကြရအောင်။

[10]:
```
!time python mm_demo.py --device 0 translate --src mya_Mymr --tgt eng_Latn␣
↪--text "မန္တလေးမှာ ပြောပြောနေကြတဲ့ တစ်ပြ ဆိုတာ ဘာလဲဗျ"
```

```
2025-09-30 19:27:31.535959: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 19:27:31.544415: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759235251.554189  454507 cuda_dnn.cc:8579] Unable to register cuDNN
```

```
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759235251.557446  454507 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759235251.565580  454507 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759235251.565602  454507 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759235251.565604  454507 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759235251.565606  454507 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 19:27:31.568044: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading translation model: facebook/nllb-200-distilled-600M (device=0)
src=mya_Mymr tgt=eng_Latn
Device set to use cuda:0

--- TRANSLATION ---

What's the one they're talking about in Mandalay?

real    0m8.472s
user    0m10.351s
sys     0m2.111s
```

အထက်မှာ မြင်ရတဲ့ အတိုင်းပါပဲ။ “တစ်ပြ” ဆိုတာကို “တစ်” လို့ တိုက်ရိုက်ယူသွားတာ ဖြစ်ဖို့ များပါတယ်။ Machine translation သုတေတနအလုပ်မှာလည်း NLP သမားတွေအနေနဲ့ မြန်မာစာသမားတွေအနေနဲ့ လုပ်စရာတွေ အများကြီးပါလို့ ပြောချင်ပါတယ်။

1.6 NER (named-entity recognition)

Name Entity Recognition (NER) လို့ခေါ် NLP အလုပ်ကတော့ ပြောလိုက်တဲ့အကြောင်းအရာ၊ ရေးထားတဲ့ စာကြောင်းတွေထဲကနေ လူနာမည်တွေ၊ မြို့နာမည်တွေ၊ အဖွဲ့အစည်းနာမည်တွေ၊ ရက်စွဲ၊ အရေအတွက် စတဲ့အပိုင်းတွေကို ခွဲခြားတဲ့အလုပ်၊ သိအောင် ကွန်ပျူတာမော်ဒယ်ကို သင်ပေးရတဲ့ အလုပ်ပါ။ မြန်မာစာအတွက်

NER အလုပ်ကိုလည်း ကျွန်တော်နဲ့ ကျွန်တော့ ကျောင်းသားတွေ လုပ်ဖြစ်နေပါတယ်။

Pretrained မော်ဒယ် တစ်ခုကို သုံးပြီး မြန်မာနဲ့ဆိုင်တဲ့ NER tag တွေကို ဘယ်လောက်ထိ လုပ်ပေးနိုင်သလဲဆိုတာကို လက်တွေ့စမ်းကြည့်ကြရအောင်ပါ။

Pretrained model link: https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl
myNER corpus link: https://github.com/ye-kyaw-thu/myNER

myNER corpus က ကျွန်တော်ရဲ့ မဟာတန်း ကျောင်းသား တစ်ယောက်ဖြစ်ခဲ့တဲ့ မောင်ကောင်းလွင်သန့် (**Assumption University, Thailand**) နဲ့ တခြား Lab အဖွဲ့ဝင်တွေ အတူတူပြင်ဆင်ခဲ့ကြတာပါ။

[14]:
```
!time python mm_demo.py --device 0 ner --text "ရန်ကုန်မြို့တွင် ရွှေငွေထွန်း ကုမ္ပဏီ ␣
↪သည် မကြာသေးမီက ရောင်းကုန်များကို ပြသခဲ့သည်"
```

```
2025-09-30 21:05:18.600542: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 21:05:18.609028: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759241118.618630  455832 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759241118.621725  455832 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759241118.629640  455832 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241118.629654  455832 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241118.629656  455832 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241118.629657  455832 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 21:05:18.631899: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
```

```
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading NER model: Davlan/xlm-roberta-base-ner-hrl (device=0)
Device set to use cuda:0
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- NAMED ENTITIES ---

{'entity_group': 'LOC', 'score': np.float32(0.99793446), 'word': ' ရန်ကုန်',
'start': 0, 'end': 7}
{'entity_group': 'ORG', 'score': np.float32(0.99985075), 'word': '', 'start':
16, 'end': 17}
{'entity_group': 'ORG', 'score': np.float32(0.9996956), 'word': ' ရွှေဂွေထွန်း
ကုမ္ပဏီ', 'start': 17, 'end': 37}

real    0m6.553s
user    0m10.429s
sys     0m0.599s
```

ဒီနေရာမှာ အမျိုးအစားခွဲခြားထားတဲ့ LOC ဆိုတာက location (တည်နေရာ) ပါ။
ပြီးတော့ ORG ဆိုတာက Organization အဖွဲ့အစည်းပါ။
မှန်မှန်ကန်ကန် ခွဲခြားပေးနိုင်ပါတယ်။

[15]:
```
!time python mm_demo.py --device 0 ner --text "ရတနာ ရွှေတောင်ကြီး ဝန်ဆောင်မှု ␣
↪ကုမ္ပဏီ လီမိတက် နားမှာ ငါ နေတယ်"
```

```
2025-09-30 21:07:25.154601: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 21:07:25.163002: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759241245.172950  455910 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759241245.176320  455910 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759241245.184609  455910 computation_placer.cc:177] computation
```

```
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241245.184620  455910 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241245.184622  455910 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241245.184623  455910 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 21:07:25.186880: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading NER model: Davlan/xlm-roberta-base-ner-hrl (device=0)
Device set to use cuda:0
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- NAMED ENTITIES ---

{'entity_group': 'ORG', 'score': np.float32(0.99064744), 'word': ' ရတနာ
ရွှေတောင်ကြီး ဝန်ဆောင်မှု ကုမ္ပဏီ', 'start': 0, 'end': 38}

real    0m6.467s
user    0m10.329s
sys     0m0.638s
```

ဒီ xlm-roberta-base-ner-hrl မော်ဒယ်က မြန်စာလုံးအတွဲ ကုမ္ပဏီနာမည်ကို မှန်မှန်ကန်ကန် ORG ဆိုပြီး သိပါတယ်။
ကောင်းပြီ၊ ဒီတခါတော့ မြန်မာမြို့၊ ရွာနာမည်တွေ ပါတဲ့ စာကြောင်းနဲ့ ထပ်စမ်းကြည့် ပါမယ်။

[16]:
```
!time python mm_demo.py --device 0 ner --text "ချောင်းဦး ရွာ သည် မန္တလေး ␣
↪တိုင်းဒေသကြီး ၊ ကျောက်ဆည် ခရိုင် ၊ တံတားဦး မြို့နယ် ၊ သူငယ်တော် ကျေးရွာ အုပ်စု ၌ ␣
↪တည်ရှိ သည် ။"
```

```
2025-09-30 21:14:47.802659: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
```

```
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 21:14:47.811046: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759241687.820967  456081 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759241687.824270  456081 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759241687.832311  456081 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241687.832322  456081 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241687.832323  456081 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759241687.832325  456081 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 21:14:47.834682: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading NER model: Davlan/xlm-roberta-base-ner-hrl (device=0)
Device set to use cuda:0
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- NAMED ENTITIES ---

{'entity_group': 'LOC', 'score': np.float32(0.9997936), 'word': '', 'start': 0,
'end': 1}
{'entity_group': 'LOC', 'score': np.float32(0.9995262), 'word': ' ချောင်းဦး
```

```
ရွာ', 'start': 0, 'end': 14}
{'entity_group': 'LOC', 'score': np.float32(0.9998418), 'word': '', 'start': 18,
'end': 19}
{'entity_group': 'LOC', 'score': np.float32(0.99923325), 'word': ' မန္တလေး ',
'start': 19, 'end': 27}
{'entity_group': 'LOC', 'score': np.float32(0.99983656), 'word': '', 'start':
42, 'end': 43}
{'entity_group': 'LOC', 'score': np.float32(0.999601), 'word': ' ကျောက်ဆည်
ခရိုင်', 'start': 43, 'end': 59}
{'entity_group': 'LOC', 'score': np.float32(0.9997385), 'word': '', 'start': 61,
'end': 62}
{'entity_group': 'LOC', 'score': np.float32(0.9805145), 'word': ' တံတားဦး
မြို့နယ်', 'start': 62, 'end': 78}
{'entity_group': 'LOC', 'score': np.float32(0.9641382), 'word': ' သူငယ်တော်
ကျေးရွာ', 'start': 80, 'end': 98}

real    0m6.700s
user    0m10.257s
sys     0m0.688s
```

အားလုံးလိုလို သိတယ်လို့ ပြောလို့ ရပါတယ်။
ကျွန်တော်တို့ myNER ကောပတ်စ်ထဲက စာကြောင်းတစ်ကြောင်းကို ယူပြီး စမ်းကြည့်ခဲ့တာပါ။ လက်တွေ့မှာ အောက်ပါလိုမျိုး Conll Format နဲ့ လေဘယ်ထိုး ကြရပါတယ်။

ချောင်းဦး n B-LOC
ရွာ n E-LOC
သည် ppm O
မန္တလေး n B-LOC
တိုင်းဒေသကြီး n E-LOC
၊ punc O
ကျောက်ဆည် n B-LOC
ခရိုင် n E-LOC
၊ punc O
တံတားဦး n B-LOC
မြို့နယ် n E-LOC
၊ punc O
သူငယ်တော် n B-LOC
ကျေးရွာ n I-LOC
အုပ်စု n E-LOC
၌ ppm O
တည်ရှိ v O
သည် ppm O
။ punc O

myNER ကောပတ်စ် ကို စိတ်ဝင်စားတဲ့သူတွေက အောက်ပါလင့်ကနေ လေ့လာလို့ရပါတယ်။
GitHub Link: https://github.com/ye-kyaw-thu/myNER

[1]:
```
!time python mm_demo.py --device 0 ner --text "၂၀၁၄ သန်းခေါင် စာရင်း အရ ␣
↪ကတ်တောင် ကျေးရွာ အုပ်စု တွင် ကျား ၂၅၆၆ ဦး ၊ မ ၂၅၇၉ ဦး ၊ လူဦးရေ စုစုပေါင်း ␣
↪၅၁၄၅ ဦး နေထိုင် သည် ။"
```

```
2025-09-30 21:33:45.870274: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 21:33:45.878545: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759242825.888359  457254 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759242825.891547  457254 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759242825.899348  457254 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759242825.899358  457254 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759242825.899360  457254 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759242825.899361  457254 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 21:33:45.901668: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading NER model: Davlan/xlm-roberta-base-ner-hrl (device=0)
Device set to use cuda:0
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`.
```

```
  return forward_call(*args, **kwargs)

--- NAMED ENTITIES ---

{'entity_group': 'LOC', 'score': np.float32(0.9976944), 'word': ' ကတ်တောင်',
'start': 24, 'end': 32}

real    0m6.563s
user    0m10.373s
sys     0m0.649s
```

ဒီတခါမှာတော့ မြန်မာ နှစ်တွေ၊ ဂဏန်းတွေကိုတော့ မခွဲခြားပေးနိုင်တာကို တွေ့ရပါတယ်။
myNER ကောပတ်စ် ထဲမှာက အောက်ပါလိုမျိုး ကျွန်တော်တို့ tagging လုပ်ထားကြပါတယ်။

၂၀၁၄ num S-DATE
သန်းခေါင် n O
စာရင်း n O
အရ ppm O
ကတ်တောင် n B-LOC
ကျေးရွာ n I-LOC
အုပ်စု n E-LOC
တွင် ppm O
ကျား n O
၂၅၆၆ num S-NUM
ဦး part O
၊ punc O
မ part O
၂၅၇၉ num S-NUM
ဦး part O
၊ punc O
လူဦးရေ n O
စုစုပေါင်း n O
၅၁၄၅ num S-NUM
ဦး part O
နေထိုင် v O
သည် ppm O
။ punc O

[3]:
```
!time python mm_demo.py --device 0 ner --text "ခင်လေးမူ သည် မြန်မာ မေကျာ ယူ ပ
↪ထား သည် ။"
```

```
2025-09-30 21:38:36.531467: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 21:38:36.539965: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
```

```
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759243116.549678  457358 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759243116.552921  457358 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759243116.560998  457358 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759243116.561009  457358 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759243116.561011  457358 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759243116.561012  457358 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 21:38:36.563246: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading NER model: Davlan/xlm-roberta-base-ner-hrl (device=0)
Device set to use cuda:0
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- NAMED ENTITIES ---

{'entity_group': 'PER', 'score': np.float32(0.757351), 'word': '', 'start': 0,
'end': 1}
{'entity_group': 'PER', 'score': np.float32(0.74629366), 'word': ' ə𝖢', 'start':
0, 'end': 3}

real    0m6.640s
```

```
user    0m10.382s
sys     0m0.688s
```

အထက်ပါအတိုင်း မြန်မာနာမည် “ခင်လေးမူ” ကိုတော့ မော်ဒယ်က NER အဖြစ် မသိနိုင်တာကို တွေ့ကြရပါတယ်။ ကျွန်တော်တို့ myNER ကောပတ်စ်ထဲမှာတော့ အောက်ပါအတိုင်း လေဘယ်ထိုးထားပါတယ်။

ခင်လေးမူ n S-PER
သည် ppm O
မြန်မာ n O
မေဂျာ n O
ယူ v O
ထား part O
သည် ppm O
။ punc O

[5]:
```
!time python mm_demo.py --device 0 ner --text "ဂျပန် နိုင်ငံ ၊ တိုကျို မြို့ ၊ ␣
↪ကုလသမဂ္ဂ တက္ကသိုလ် ဌာနချုပ် United Nations University UNU တွင် ဦးသန့် ဂုဏ်ပြု ␣
↪သင်ကြား ပို့ချ ချက် များ ဟူသော သင်ခန်းစာ ကို ပုံမှန် သင်ကြား ပေး လျက် ရှိ သည် ။"
```

```
2025-09-30 21:46:40.966473: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 21:46:40.975054: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759243600.984892  457558 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759243600.988164  457558 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759243600.996274  457558 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759243600.996285  457558 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759243600.996287  457558 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759243600.996288  457558 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 21:46:40.998534: I tensorflow/core/platform/cpu_feature_guard.cc:210]
```

```
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
Loading NER model: Davlan/xlm-roberta-base-ner-hrl (device=0)
Device set to use cuda:0
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- NAMED ENTITIES ---

{'entity_group': 'LOC', 'score': np.float32(0.9998431), 'word': '', 'start': 0,
'end': 1}
{'entity_group': 'LOC', 'score': np.float32(0.83837444), 'word': ' ဂျပန်
နိုင်ငံ', 'start': 0, 'end': 13}
{'entity_group': 'LOC', 'score': np.float32(0.9991511), 'word': '', 'start': 15,
'end': 16}
{'entity_group': 'LOC', 'score': np.float32(0.9966246), 'word': ' တိုကျို မြို့',
'start': 16, 'end': 29}
{'entity_group': 'ORG', 'score': np.float32(0.959226), 'word': ' ကုလသမဂ္ဂ ',
'start': 31, 'end': 41}
{'entity_group': 'ORG', 'score': np.float32(0.9708598), 'word': 'United Nations
University UNU', 'start': 59, 'end': 89}

real    0m6.530s
user    0m10.343s
sys     0m0.658s
```

အထက်ပါ အဖြေကနေ လေ့လာသိရှိနိုင်တာက လက်ရှိသုံးနေတဲ့ မော်ဒယ်က မြန်မာလူပုဂ္ဂိုလ်နာမည် ဖြစ်တဲ့ “ဦးသန့်” နဲ့ “ဌာနချုပ်” ကိုတော့ NER အနေနဲ့ မခွဲခြားပေးနိုင်၊ မသိနိုင်သေးဘူး ဆိုတဲ့ အချက်ပါပဲ။

ဂျပန် n B-LOC
နိုင်ငံ n E-LOC
၊ punc O
တိုကျို n B-LOC
မြို့ n E-LOC
၊ punc O
ကုလသမဂ္ဂ n B-ORG
တက္ကသိုလ် n I-ORG
ဌာနချုပ် n E-ORG

United fw O
Nations fw O
University fw O
UNU fw O
တွင် ppm O
ဦးသန့် n S-PER
ဂုဏ်ပြု v O
သင်ကြား v O
ပို့ချ v O
ချက် n O
များ part O
ဟူသော part O
သင်ခန်းစာ n O
ကို ppm O
ပုံမှန် adv O
သင်ကြား v O
ပေး part O
လျက် conj O
ရှိ v O
သည် ppm O
။ punc O

1.7 Sentence similarity (paraphrase detector)

ဒီနေရာမှာတော့ paraphrase ဟုတ်သလား၊ မဟုတ်ဘူးလား ဆိုတာကို multilingual မော်ဒယ်တစ်ခုကို သုံးပြီး လက်တွေ့ ခွဲခြားပြပါမယ်။
Paraphrase ဆိုတာကတော့ စာကြောင်း နှစ်ကြောင်းက ရေးထားတဲ့ပုံစံ၊ သုံးထားတဲ့ စာလုံးတွေ အတိအကျမတူပေမဲ့ အဓိပ္ပါယ်အားဖြင့် တူတဲ့ စာကြောင်းတွေပါ။

ဘာသာဗေဒ ပညာရှင်တွေကတော့ ပိုသိကြမှာပါ။ ကျွန်တော်တို့ လူတွေက အခြေအနေ အချိန်အခါ၊ နားထောင်မယ့် တဖက်လူရဲ့ ရာထူး၊ အဆင့်အတန်း စတာတွေပေါ်ကို မူတည်ပြီး စကားကို ပုံစံမျိုးစုံနဲ့ ပြောကြပါတယ်။ အရေးအသားမှာလည်း ထိုနည်းလည်းကောင်းပါပဲ။ အဲဒီ အပိုင်းတွေကို ကွန်ပျူတာ မော်ဒယ်က ထဲထဲဝင်ဝင် သိမှသာ ကွန်ပျူတာနဲ့ လူနဲ့အကြားရဲ့ ဆက်ဆံရေး၊ အတူတွဲ အလုပ်လုပ်ကြတဲ့အပိုင်းတွေက အဆင်ပြေကြမှာမို့၊ လက်တွေ့ အသုံးဝင်တဲ့ ဆားဗစ်မျိုးကို ထောက်ပံ့ပေးနိုင်မှာမို့ အဲဒီ paraphrase, sentence similarity အပိုင်းကိုလည်း NLP R&D အနေနဲ့ AI ရဲ့ အစိတ်အပိုင်း တခုအနေနဲ့ လုပ်ကြရပါတယ်။

Model link: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

myParaphrase GitHub Link: https://github.com/ye-kyaw-thu/myParaphrase

[6]:
```
! time python mm_demo.py --device 0 sim --text1 "ဒီ နေ့ အလုပ် လုပ် မှာ လား"␣
↪--text2 "ဒီ နေ့ အလုပ် လာ မှာ လား"
```

```
2025-09-30 22:06:58.973512: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
```

```
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 22:06:58.981847: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759244818.991604  457884 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759244818.994874  457884 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759244819.002914  457884 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759244819.002926  457884 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759244819.002928  457884 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759244819.002929  457884 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 22:06:59.005277: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `BertSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- SIMILARITY ---

Score: 0.9362

real    0m9.085s
user    0m9.844s
sys     0m0.689s
```

မော်ဒယ်ရဲ့ ဆုံးဖြတ်ချက်က မှားပါတယ်။
ဘာကြောင့်လဲ ဆိုတော့ အထက်မှာ မြင်ရတဲ့အတိုင်း သူပေးထားတဲ့ အမှတ်က ၁ နဲ့ အရမ်းကပ်နေလို့ပါ။

[8]:
```
! time python mm_demo.py sim --text1 "ဒီ မှာ ငါ့ နံပါတ် ။" --text2 "ဒီ မှာ သူ့
↪နံပါတ် ။"
```

```
2025-09-30 22:11:31.116821: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 22:11:31.125310: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759245091.135173  458069 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759245091.138476  458069 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759245091.146716  458069 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759245091.146726  458069 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759245091.146727  458069 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759245091.146729  458069 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 22:11:31.149083: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `BertSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)
```

```
--- SIMILARITY ---

Score: 0.9411

real    0m9.181s
user    0m9.822s
sys     0m0.716s
```

“ဒီ မှာ ငါ့ နံပါတ် ။” ဆိုတဲ့ စာကြောင်းနဲ့ “ဒီ မှာ သူ့ နံပါတ် ။” ဆိုတဲ့ စာကြောင်း နှစ်ကြောင်းမှာ ဆိုရင် တကယ်တမ်း “ငါ့” နဲ့ “သူ့” ဆိုတဲ့ စကားလုံး နှစ်လုံးပဲ ကွာခြားတာမို့လို့ ပုံမှန် string similarity တွက်တဲ့နည်းတွေနဲ့ တွက်ရင် sentence similarity က တူမှာပါပဲ။ ဒါပေမဲ့ paraphrase အနေနဲ့ ကြည့်ရင်တော့ အဓိပ္ပါယ်က လုံးဝ ဆန့်ကျင်ဘက်ပါ။ ဒီလိုမျိုး စာကြောင်းမျိုးတွေက မော်ဒယ်အတွက်က ကွဲကွဲပြားပြား ခွဲခြားပေးနိုင်ဖို့ အတော်ကို ခက်ပါတယ်။ မြန်မာစာနဲ့ သက်ဆိုင်တဲ့ Training data ကို သေချာပြင်ဆင်ပေးထားမှပဲ လက်ရှိလို အမှားမျိုးတွေကို မော်ဒယ်က ရှောင်နိုင်မယ်လို့ နားလည်ပါတယ်။

[9]:
```
! time python mm_demo.py sim --text1 "ဆူးလေ ဘုရား အလွန် မှာ ။" --text2 "␣
↪ဆူးလေ ဘုရား နား လေး မှာ ။"
```

```
2025-09-30 22:21:56.489101: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 22:21:56.497295: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759245716.506873  458239 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759245716.510068  458239 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759245716.517911  458239 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759245716.517922  458239 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759245716.517924  458239 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759245716.517926  458239 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
```

```
2025-09-30 22:21:56.520257: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
/home/ye/.local/lib/python3.12/site-packages/torch/nn/modules/module.py:1784:
FutureWarning: `encoder_attention_mask` is deprecated and will be removed in
version 4.55.0 for `BertSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)

--- SIMILARITY ---

Score: 0.9938

real    0m9.156s
user    0m9.801s
sys     0m0.729s
```

ဗမာစကား၊ မြန်မာစကား ကောင်းကောင်း နားလည်တဲ့ သူဆိုရင် “ဆူးလေ ဘုရား အလွန် မှာ ။” ဆိုတဲ့ စာကြောင်းနဲ့ “ဆူးလေ ဘုရား နား လေး မှာ ။” ဆိုတဲ့ စာကြောင်းနှစ်ကြောင်းက အဓိပ္ပါယ် မတူတာကို အလွယ်တကူ သိနိုင်ပေမဲ့ ကွန်ပျူတာမော်ဒယ် အနေနဲ့က ခက်ပါတယ်။ အထက်မှာ မြင်ရတဲ့ ရလဒ်အတိုင်းပါပဲ မော်ဒယ်က paraphrase ဖြစ်နိုင်ခြေက 0.9938 လို မှားဆုံးဖြတ်ထားပါတယ်။

ကျွန်တော်တို့ LU Lab. က ပြင်ဆင်ထားတဲ့ **myParaphrase** ကောပတ်စ်မှာက စာကြောင်းရေအတွဲပေါင်း လေးသောင်းကျော် ပြင်ဆင်ပြီး ရဲ့ပေးထားပါတယ်။ ကျွန်တော်ရဲ့ ဒေါက်တာတန်းကျောင်းသူ တစ်ယောက်ဖြစ်ခဲ့တဲ့ မြင့်မြင့်ဌေး **UTYCC (University of Technology - Yatanarpon Cyber City)** နဲ့ အတူတူ ပြင်ဆင်ခဲ့တဲ့ ဒေတာပါ။

1.8 Spell correction

1.8.1 (requires a dict file: wrong<TAB>correct OR term<TAB>freq)

နောက်ဆုံးအနေနဲ့ ဒီမိုလုပ်ပြချင်တာကတော့ ဘယ်တော့မှ ၁၀၀% နှုန်းမှန်ကန်တဲ့ ရလဒ်ကို မရနိုင်တဲ့ NLP သုတေသန ဖြစ်တဲ့ စာလုံးပေါင်းအမှားတွေကို စစ်တဲ့၊ ပြင်တဲ့ အလုပ်ပါပဲ။

သိစေချင်တာက ဒီမော်ဒယ်ကတော့ LLM သို့မဟုတ် GPT မော်ဒယ် မဟုတ်ပါဘူး။
Statistical model တစ်ခုဖြစ်တဲ့ SymSpell ဆိုတဲ့ မော်ဒယ်ကိုပဲ သုံးပြထားပါတယ်။

SymSpell GitHub Link: https://github.com/wolfgarbe/SymSpell

ဇော်ဂျီနဲ့ဖောင့် ရိုက်ထားတဲ့ “မဂ□လာ” ဆိုတာကိုပဲ စစ်ခိုင်းကြည့်ရအောင်ပါ။

[10]:
```
!time python mm_demo.py spellcheck     --word "မဂ□လာ"     --dict ./data/g2p.
 ↪freq     --mode fuzzy     --max-edit-distance 5     --num-suggestions 10
```

```
2025-09-30 22:54:41.200888: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-09-30 22:54:41.209389: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759247681.219213  458776 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759247681.222518  458776 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759247681.230815  458776 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759247681.230825  458776 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759247681.230827  458776 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759247681.230828  458776 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-09-30 22:54:41.233195: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

--- SPELLCHECK ---

ဖဂ□လ -> ဖဆလ (distance=2, freq=1)
ဖဂ□လ -> ဖဇလ (distance=2, freq=1)
ဖဂ□လ -> ဖလ (distance=2, freq=1)

real    0m4.768s
user    0m9.451s
sys     0m0.396s
```

တကယ်က စာလုံးပေါင်းအမှားကို စစ်ဖို့အတွက် ကျွန်တော်ပြင်ပေးထားတဲ့ အဘိဓာန်ထဲမှာ "မင်္ဂလာ" ဆိုတဲ့ စာလုံးက ပါပါတယ်။ ဒါပေမဲ့ အဘိဓာန်ပဲသုံးတဲ့ မော်ဒယ်အနေနဲ့က ဆွဲထုတ်ပေးနိုင်ဖို့ ခက်ခဲတာပါ။

အလွယ်ရှင်းပြရရင်တော့ SymSpell ဆိုတဲ့ မော်ဒယ်က မှားနေတဲ့ စာလုံးနဲ့ အနီးစပ်ဆုံး တူညီနိုင်တဲ့ စာလုံးကို အဘိဓာန်ထဲကနေ ဆွဲထုတ်ယူဖို့အတွက် edit distance ဆိုတဲ့ တန်ဖိုးနဲ့ တိုင်းတာပါတယ်။ Edit distance က သုည ဆိုရင် စာလုံးနှစ်လုံးက တူလို့ပါ။ Edit distance ရဲ့ တန်ဖိုးက ကြီးလာတာနဲ့အမျှ စာလုံးနှစ်လုံးက ဘယ်လောက် မတူဘူးလဲ ဆိုတာကို ပြသတာမို့ပါ။ လက်တွေ့ စာလုံးပေါင်းအမှားတွေက ပုံစံမျိုးစုံနဲ့ (ဥပမာ အသံဆင်တူလို့မှားတာ၊ encoding လို့ခေါ်တဲ့ စာလုံးတွေကို ကွန်ပျူတာမှာ သိမ်းဖို့သုံးတဲ့ နည်းမတူလို့ မပြပေးနိုင်တာ၊ စာရိုက်တဲ့ ကီးဘုတ်၊ ဖုန်း လက်ကွက်နီးလို့ မှားရိုက်မိတာ၊ Facebook မှာ ပြောချင်တာကို သွယ်ဝိုက်ပြီးပြောတာ စသည်ဖြင့်) ရှိပါတယ်။ Edit distance တစ်ခုတည်းနဲ့ တိုင်းတာပြီး ဆွဲထုတ်လို့ မရတဲ့ စာလုံးတွေလည်း ရှိနိုင်လို့ပါ။

ဒီနေရာမှာတော့ Edit distance တွက်တဲ့အပိုင်းကို အကျယ်မရှင်းတော့ပါဘူး။ စိတ်ဝင်စားတဲ့ သူတွေက အောက်ပါ လင့်ကနေလည်း လေ့လာလို့ ရပါတယ်။

Wiki Link: https://en.wikipedia.org/wiki/Edit_distance

မြန်မာစာလုံးတွေ စာကြောင်းတွေရဲ့ string similarity ကိုပဲ ဦးတည်ပြီး တိုင်းတာတဲ့ ပရိုဗိုဇယ် တစ်ခုကိုလည်း မဟာတန်း သုတေသနအနေနဲ့ ခိုင်ဆုဝေ (လက်ရှိမှာ Akita University မှာ Assistant Professor) နဲ့အတူ တွဲလုပ်ဖြစ်ခဲ့ကြပါသေးတယ်။

Paper Link: https://aclanthology.org/2019.nsurl-1.14/

1.9 Dictionary of SymSpell

SymSpell အတွက်က အဘိဓာန် နှစ်မျိုးကို ပြင်ဆင်ကြရပါတယ်။

တစ်မျိုးက အမှားစာလုံးအမှန်စာလုံး ဆိုတဲ့ ပုံစံနဲ့ စာလုံး တစ်လုံးချင်းစီ ရိုက်ထည့်ထားတဲ့ အဘိဓာန်ပါ။ ဥပမာ အောက်ပါလိုမျိုး

[11]:
```
! cat ./data/spell_pair.txt
```

```
မဂ္ဂလာ	မင်္ဂလာ
နေကောလား	နေကောင်းလား
ဘာဖစ်ဖစ်	ဘာဖြစ်ဖြစ်
```

နောက်ပုံစံတစ်မျိုးကတော့ စာလုံးကြိမ်နှုန်း ဆိုတဲ့ ပုံစံပါ။
ဒီနေရာမှာ ပြောတဲ့ ကြိမ်နှုန်း (frequency) ဆိုတာက ကောပတ်စ်ထဲမှာ စာလုံးတစ်လုံးချင်းစီက ဘယ်နှခါ ပါသလဲ၊ ဘယ်နှခါ သုံးထားသလဲ ဆိုတာကို ရေတွက်ထားတဲ့ တန်ဖိုးကို ဆိုလိုတာပါ။ အဲဒီလို စာလုံး တစ်လုံးချင်းစီရဲ့ တကယ့်မှန်ကန်တဲ့၊ တန်ဖိုးရှိတဲ့ ကြိမ်နှုန်းကို ရယူနိုင်ဖို့အတွက်က စာကြောင်းရေ များနိုင်သမျှ များတဲ့ ကောပတ်စ်က လိုအပ်ပါတယ်။ ပြီးတော့ စာလုံးတွေကို မှန်မှန်ကန်ကန် ဖြတ်ထားတဲ့ အပိုင်းကလည်း အရေးကြီးပါတယ်။ အဲဒီအတွက်လည်း ကျွန်တော်တို့ Lab က ပြင်ဆင်နေတာတွေရှိပါတယ်။

ဒီ notebook မှာ စာလုံးပေါင်းအမှား ရှာတဲ့ ဥပမာအတွက်ကတော့ စာလုံးအားလုံးကို ကြိမ်နှုန်းအားလုံး အတူတူထားထားခဲ့ပါတယ်။ ဆိုလိုတာက စာလုံးအားလုံးကို ညီမျှတဲ့ တန်ဖိုးပေးထားတာပါ။ အင်္ဂလိပ်လိုက equal weight ပေးထားတယ်လို့ခေါ်ပါတယ်။ အောက်ပါအတိုင်းပါ။

[12]:
```
! head ./data/g2p.freq
```

ကကတစ် 1
ကကတိုး 1
ကကုသန် 1
ကကုသန် 1
ကကူရံ 1
ကကြိုး 1
ကကြိုးတန်ဆာ 1
ကကြီကကြောင်လုပ် 1
ကကြီး 1
ကကြီးထွန် 1

[13]:
```
! shuf ./data/g2p.freq | head -n 30
```

ခြေတစ်ပေါင်ကျိုး 1
ဆွဲခြင်း 1
ခရုစုတ်အနက် 1
တုံးပေကတ်သတ်ခံ 1
မြင်းခံ 1
ကာသာပေး 1
ငရုံပတူ 1
အုပ်ဆောင်း 1
အခေါင် 1
ကြိတ်ချေ 1
သည်သို့ 1
သန်းခေါင်ယံ 1
ငေါက်တောက် 1
ရင်ခွင် 1
ခန့်မှန်း 1
နားနားနေနေ 1
ဦးယဉ်းနာ 1
ဘဝင်ကျ 1
အငြင်း 1
အကြမ်းဖျင်း 1
မိုးခြိမ်း 1
ရင်ကြားစေ့ 1
ဘန်းမုန့် 1
တင်ကူး 1
ရာပြတ် 1
ကုတ်ကုတ်ကတ်ကတ် 1
မိုးယံ 1
ဆိုးမွေ 1
ဂါထာရွတ် 1

```
ဖိအား    1
shuf: write error: Broken pipe
```

မြန်မာစာအဖွဲ့က ထုတ်ဝေထားတဲ့ မြန်မာအဘိဓာန်ထဲက စာလုံးတွေအပြင် ထပ်ဖြည့်ရိုက်ထားတဲ့ အဘိဓာန်ပါ။ ပြင်ဖြစ်သွားခဲ့တာက ကျွန်တော် ဂျပန်နိုင်ငံ၊ ကျိုတိုမြို့က NICT (National Institute of Information and Communications Technology) မှာ အလုပ်လုပ်ခဲ့စဉ်က ပရောဂျက်တစ်ခု ဖြစ်တဲ့ VoiceTra ပရောဂျက်အတွက် လိုအပ်တဲ့ grapheme-to-phoneme ကို ပြောင်းလဲ ပေးတဲ့ သုတေသန အလုပ်တွေအတွက်ပါ။ အဲဒီထဲ g2p အဘိဓာန်ထဲက စာလုံးတွေကို ယူထားတာပါ။ လက်ရှိ ဗားရှင်းမှာက စာလုံးရေ နှစ်သောင်းလေးထောင်ကျော် ရှိပါတယ်။

[14]:
```
!wc ./data/g2p.freq
```

```
 24798  49596 741228 ./data/g2p.freq
```

g2p အဘိဓာန်ကို သုံးပြီး spelling suggestion ကို လုပ်ခိုင်းကြည့်ကြရအောင်။ ဆွဲထုတ်ပေးနိုင်တာကော၊ ဆွဲမထုတ်ပေးနိုင်တာကော အမျိုးမျိုး တွေ့ကြရပါလိမ့်မယ်။

[15]:
```
! time python mm_demo.py spellcheck     --word "မင်ဂလာ"     --dict ./data/g2p.
 ↪freq     --mode fuzzy     --max-edit-distance 5     --num-suggestions 10
```

```
2025-10-01 00:45:49.584477: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-10-01 00:45:49.592698: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759254349.602125  460528 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759254349.605445  460528 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759254349.613267  460528 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254349.613279  460528 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254349.613280  460528 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254349.613282  460528 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
```

```
2025-10-01 00:45:49.615595: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

--- SPELLCHECK ---

မင်ဂလာ -> မင်္ဂလာ (distance=1, freq=1)

real    0m4.918s
user    0m9.452s
sys     0m0.514s
```

အထက်မှာ မြင်ရတဲ့အတိုင်းပါပဲ “မင်ဂလာ” ဆိုတဲ့ စာလုံးပေါင်းအမှားအတွက် အဖြစ်နိုင်ဆုံး သို့မဟုတ် မှန်တဲ့ စာလုံးကိုတော့ SymSpell မော်ဒယ်က ဆွဲထုတ်ပေးနိုင်ပါတယ်။

[16]:
```
! time python mm_demo.py spellcheck     --word "မလ်ာ"     --dict ./data/g2p.
  ↪freq     --mode fuzzy     --max-edit-distance 5     --num-suggestions 10
```

```
2025-10-01 00:49:03.269323: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-10-01 00:49:03.277874: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759254543.287726  460631 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759254543.291048  460631 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759254543.299114  460631 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254543.299125  460631 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254543.299127  460631 computation_placer.cc:177] computation
```

```
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254543.299128  460631 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-10-01 00:49:03.301414: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

--- SPELLCHECK ---

 မလ်ာ -> မဂ််လာ (distance=1, freq=1)

real    0m4.772s
user    0m9.409s
sys     0m0.438s
```

အထက်မှာ မြင်ရတဲ့အတိုင်းပါပဲ။ မဂ််လာ ကို ရိုက်တဲ့အခါမှာ ဂငယ် စာလုံးတစ်လုံးတည်း ကျန်ခဲ့တာမျိုးကိုလည်း မော်ဒယ်က အမှန်ပြင်ပေးနိုင်ပါတယ်။

[17]:
```
! time python mm_demo.py spellcheck     --word "လူဆိုးးး"     --dict ./data/
 ↪g2p.freq     --mode fuzzy     --max-edit-distance 5     --num-suggestions 10
```

```
2025-10-01 00:51:02.076568: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-10-01 00:51:02.084917: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1759254662.094704  460708 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759254662.098011  460708 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759254662.106005  460708 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
```

```
target more than once.
W0000 00:00:1759254662.106015  460708 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254662.106017  460708 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254662.106019  460708 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-10-01 00:51:02.108494: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

--- SPELLCHECK ---

လူဆိုးးး -> လူဆိုး (distance=2, freq=1)

real    0m4.744s
user    0m9.399s
sys     0m0.425s
```

သုံးတဲ့ text editor ပေါ်မူတည်ပြီး ကွဲကွဲပြားပြား မြင်ရတာမျိုးမရှိတဲ့ “ညလေး” နဲ့ “အက္ခရာ ဥ” မှားရိုက်တဲ့ အမှားမျိုးကို spelling correction လုပ်ခိုင်းကြည့်ကြရအောင်။

ညလေး = ဥ (လေယာဉ်ပျံ)
အက္ခရာ ဥ = ဥ (လေယာဉ်ပျံ)

[18]:
```
! time python mm_demo.py spellcheck     --word "လေယာဉ်ပျံ"     --dict ./data/
 ↪g2p.freq     --mode fuzzy     --max-edit-distance 5     --num-suggestions 10
```

```
2025-10-01 00:55:08.129474: I tensorflow/core/util/port.cc:153] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-10-01 00:55:08.137708: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
```

```
E0000 00:00:1759254908.147178  460814 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1759254908.150389  460814 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1759254908.158343  460814 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254908.158354  460814 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254908.158356  460814 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1759254908.158357  460814 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-10-01 00:55:08.160568: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

--- SPELLCHECK ---

လေယာဉ်ပျံ -> လေယာဉ်ပျံ (distance=1, freq=1)

real    0m4.745s
user    0m9.432s
sys     0m0.401s
```

လက်တွေ့ မြန်မာစာလုံးတွေ၊ မြန်မာစာရိုက်ထည့်ထားတဲ့ဖိုင်တွေနဲ့ အလုပ်လုပ်ရတဲ့အခါမှာ terminal မှာ လုပ်ရတာမို့လို့ ဖတ်လို့မရပါဘူး။ နောက်ပြီး ဆာဗာပေါ်မှာ လုပ်ကြရတာက များတာမို့လို့ စစ်ချင်တဲ့ မြန်မာစာကြောင်းတွေကို ကော်ပီကူး ကိုယ့် လိုကယ်ကွန်ပျူတာရဲ့ text editor တစ်ခုခုမှာ ဝင်ကြည့်တာမျိုး လုပ်ကြရပါတယ်။ ပြီးတော့ အောက်ပါလိုမျိုး Unicode နံပါတ်တွေကို ရိုက်ထုတ်ပေးတဲ့ ပရိုဂရမ်တွေကို သုံးပြီး အသေးစိတ် နှိုင်းယှဉ်ကြည့်တာမျိုးလည်း လုပ်ကြရပါတယ်။

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-

"""
```

```
For printing code points (decimal, unicode) of each Myanmar character and counting total no. o
Written by Ye Kyaw Thu, Visiting Professor,
Language Semantic Technology Research Team (LST), NECTEC, Thailand

How to run:
    python print-codepoint.py --input filename
    python print-codepoint.py "word1" "word2"
    python print-codepoint.py "လေယာဉ်ပျံ" "လေယာဉ်ပျံ"
"""

import argparse
import sys
import os

def print_codepoints(text, output_file=None):
    """
    Print code points for each character in the text
    """
    if output_file:
        output_file.write(f"{text}\n")
        print(text)
    else:
        print(text)

    # Remove newline for processing but keep original for display
    processed_text = text.rstrip('\n\r')

    results = []
    for char in processed_text:
        decimal_code = ord(char)
        unicode_hex = f"U{decimal_code:04X}"
        results.append(f"{char} ({decimal_code}, {unicode_hex})")

    output_line = " ".join(results) + f", no. of char = {len(processed_text)}"

    if output_file:
        output_file.write(output_line + "\n")
        output_file.flush()
    print(output_line)

def process_file(input_file, output_file=None):
    """
    Process an input file line by line
    """
    try:
        with open(input_file, 'r', encoding='utf-8') as f:
            for line in f:
                print_codepoints(line, output_file)
```

```
                if output_file:
                    output_file.write("\n")
                print()
    except FileNotFoundError:
        print(f"Error: Input file '{input_file}' not found.", file=sys.stderr)
        sys.exit(1)
    except Exception as e:
        print(f"Error reading file: {e}", file=sys.stderr)
        sys.exit(1)

def process_words(words, output_file=None):
    """
    Process words provided as command line arguments
    """
    for word in words:
        print_codepoints(word, output_file)
        if output_file and word != words[-1]:
            output_file.write("\n")
        if word != words[-1]:
            print()

def main():
    parser = argparse.ArgumentParser(
        description='Print code points (decimal, unicode) of each Myanmar character and count t
        epilog='Examples:\n'
               '  python print-codepoint.py --input pair.txt\n'
               '  python print-codepoint.py "word1" "word2"\n'
               '  python print-codepoint.py "လေယာဉ်ပျံ" "လေယာဉ်ပျံ"\n'
               '  python print-codepoint.py --input pair.txt --output result.txt',
        formatter_class=argparse.RawDescriptionHelpFormatter
    )

    parser.add_argument(
        'words',
        nargs='*',
        help='Words to process (optional if --input is provided)'
    )

    parser.add_argument(
        '--input', '-i',
        help='Input file containing text to process'
    )

    parser.add_argument(
        '--output', '-o',
        help='Output file (default: stdout)'
    )
```

```
    args = parser.parse_args()

    # Validate arguments
    if not args.input and not args.words:
        parser.print_help()
        print("\nError: Either provide an input file with --input or words as arguments", file=
        sys.exit(1)

    if args.input and args.words:
        print("Warning: Both input file and words provided. Processing input file only.", file=

    # Setup output
    output_file = None
    if args.output:
        try:
            output_file = open(args.output, 'w', encoding='utf-8')
        except Exception as e:
            print(f"Error creating output file: {e}", file=sys.stderr)
            sys.exit(1)

    try:
        # Process input
        if args.input:
            process_file(args.input, output_file)
        else:
            process_words(args.words, output_file)

    finally:
        # Close output file if opened
        if output_file:
            output_file.close()
            print(f"Output saved to: {args.output}")

if __name__ == "__main__":
    main()
```

print-codepoint.py လို့နာမည်ပေးထားတဲ့ ပိုင်သွန် ပရိဂရမ်ကိုသုံးပြီး အမှန် (ဥလေး)၊ အမှား (အက္ခရာ ဥ) သုံးပြီး ရိုက်ထားတာကို နှိုင်းယှဉ်ကြည့်ရအောင်။

[20]:
```
! python ./print-codepoint.py " လေယာဉ်ပျံ" " လေယာဥ်ပျံ"
```

```
လေယာဉ်ပျံ
လ (4124, U101C) ေ (4145, U1031) ယ (4122, U101A) ာ (4140, U102C) ဉ (4105, U1009)
် (4154, U103A) ပ (4117, U1015) ျ (4155, U103B) ံ (4150, U1036), no. of char = 9

လေယာဥ်ပျံ
```

```
လ (4124, U101C) ◌ေ (4145, U1031) ယ (4122, U101A) ◌ာ (4140, U102C) ဥ (4133,␣
↪U1025)
◌် (4154, U103A) ပ (4117, U1015) ◌ျ (4155, U103B) ◌ံ (4150, U1036), no. of char␣
↪= 9
```

လက်တွေ့ အမှားတွေကို ပြင်ပေးနိုင်ဖို့က အောက်ပါလိုမျိုး အမှား|||အမှန် အတွဲတွေ ပြင်ထားတဲ့ အဘိဓာန်တွေကို စာလုံး တစ်လုံးချင်းစီအတွက်သာမကပဲ စာကြောင်း အနေနဲ့ပါ ရှာဖွေတာ၊ စုဆောင်းတာ၊ လက်နဲ့ အမှန်ပြင်ထားတာတွေကို လုပ်ကြရပါတယ်။

အောက်ပါ စာကြောင်း ဥပမာ တချို့က လက်ရှိ အချိန်ရရင်ရသလို Facebook ရဲ့ comment တွေမှာ ရိုက်ထားတဲ့ စာလုံးပေါင်းအမှားတွေကို စုထားပြီး၊ ကိုယ်တိုင် ဉာဏ်မီသလောက် အမှန်ပြင်ထားတာတွေပါ။

မြင်ကြရတဲ့အတိုင်းပါပဲ emoji စာလုံးတွေလည်း ပါတတ်ပါတယ်။

အရည်းခြင်ရှိသော ဘဲကြီးပါ|||အရည်းအချင်းရှိသော ဘဲကြီးပါ
မသေသေးဘူးလား နှယောက်လုံးက|||မသေသေးဘူးလား နှစ်ယောက်လုံးက
#နှလုံးကိုဖိုပြီး_မအိုက်ကြပါနဲ့|||#နှလုံးကိုဖိုပြီး_မအိပ်ကြပါနဲ့
ခလေးတွေကိုစာရေစာဂုံးရေးခိုင်ရမယ် |||ခလေးတွေကိုစာစီစာကုံးရေးခိုင်းရမယ်
မှေင်တာကမီးဖြတ်ထားလို့ပါသားရယ် |||မှောင်တာကမီးဖြတ်ထားလို့ပါသားရယ်
တိတ်စမ်း တေချင်းဆိုးလေး |||တိတ်စမ်း သေခြင်းဆိုးလေး
ဟားးးး|||ဟား
မူးမူးနက် စမ်းတဝါးဝါးသွားနေကျဟာကို ခုမှဟင်းးးးးး တိတ်စမ်း |||မူးမူးနဲ့ စမ်းတဝါးဝါးသွားနေကျဟာကို ခုမှဟင်း တိတ်စမ်း
ဟိုကောင်ကဆရာကြီးလေ အမှတ်နဲ့ဆုံးတာဘဲဘာမှတ်ဖြစ်ဘူး||| ဟိုကောင်ကဆရာကြီးလေ အမှတ်နဲ့ရှုံးတာပဲဘာမှမဖြစ်ဘူး
ကျဝ်ထက်ပြန်လာမယ် -သွေ-းလင်းထက်|||ကျင့်ထပ်ပြန်လာမယ် -သွေးလင်းထက်
Osaka ရဲ့ လည်စရာတခုဖြစ်တဲ့ (မှာ မီးလောင်မှု ဖြစ်ပွါးပြီး မီးသ-တ်ကားအစီး ၇၀ နဲ့ ၃နာရီလောက် ငြိမ်းသက်ခဲ့ရပါတယ်|||Osaka ရဲ့ လည်စရာတစ်ခုဖြစ်တဲ့ (မှာ မီးလောင်မှု ဖြစ်ပွားပြီး မီးသတ်ကားအစီး ၇၀ နဲ့ ၃နာရီလောက် ငြိမ်းသတ်ခဲ့ရပါတယ်
မီးသ-တ်တပ်ဖွဲ့ဝင် နှစ်ဦး (အသက် ၅၅ နှင့် ၂၂) ◌ေ-သဆုံးခဲ့|||မီးသတ်တပ်ဖွဲ့ဝင် နှစ်ဦး (အသက် ၅၅ နှင့် ၂၂) သေဆုံးခဲ့
နောင်ဘဝမှာ အေးဂျမ်းတဲ့ နိုင်ငံသစ်မှာ မွေးဖွားလာကြပါစေ |||နောင်ဘဝမှာ အေးချမ်းတဲ့ နိုင်ငံသစ်မှာ မွေးဖွားလာကြပါစေ
အဲ့ဒီထဲမှာ မြနိမာ Region ရော ပါဝီလား |||အဲ့ဒီထဲမှာ မြန်မာ Region ရော ပါဝီလား

ကျွန်တော့် အနေနဲ့ မြန်မာစာ စာလုံးပေါင်းအမှားတွေကို ရှာဖွေတာ၊ ကွန်ပျူတာမော်ဒယ်တွေနဲ့ ပြင်နိုင်တဲ့ နည်းလမ်းတွေနဲ့ ပတ်သက်ပြီးလည်း သုတေသန ဆက်တိုက် လုပ်ဖြစ်ပါတယ်။

International conference စာတမ်း၊ ဂျာနယ်တွေကိုလည်း ကျောင်းသားတွေနဲ့ ရေးထားတာတွေ ရှိပါတယ်။

- Ei Thandar Phyu, Ye Kyaw Thu, Thazin Myint Oo, Hutchatai Chanlekha, Thepchai Supnith, “Myanmar Spelling Error Classification: An Empirical Study of Tsetlin Machine Techniques”, Journal of Intelligent Informatics and Smart Technology, Vol 10, October, 2024, pp. 1-11. (Submitted January 30, 2024; accepted July 3, 2024; revised August 24, 2024; published online November 9, 2024)

- Ei Thandar Phyu, Ye Kyaw Thu, Hutchatai Chanlekha, Kotaro Funakoshi and Thepchai Supnithi, “Exploring the Impact of Error Type Features Integration on Transformer-Based Myanmar Spelling Correction”, the 21st International Joint Conference on Computer Science and Software Engineering (JCSSE 2024), June 19-22, Phuket, Thailand, pp. 374-381
- Ei Phyu Phyu Mon, Ye Kyaw Thu, Thida San, Zun Hlaing Moe, Hnin Aye Thant, “Automatic Rule Extraction for Detecting and Correcting Burmese Spelling Errors”, The 4th ONA Conference, 17-18 December, Ministry of Posts and Telecommunications, Phnom Penh, Cambodia

အထက်ပါ စာတမ်းတွေမှာ သုံးခဲ့တဲ့ Spelling Checking Corpus (mySpell) ကိုလည်း စိတ်ဝင်စားတဲ့ သူတွေ နောက်ဆက်တွဲ သုတေသနအလုပ်တွေ ဆက်လုပ်နိုင်ကြအောင် ကျွန်တော့ Github နဲ့ LU Lab HuggingFace တွေမှာ ရှဲပေးနိုင်ဖို့ ပြင်ဆင်နေပါတယ်။

[ ]: