syllable_word_tokenization_demo

October 1, 2025

1 Syllable and Word Tokenization Demo for Myanmar Language

Purpose: This notebook demonstrates how to perform syllable-unit and word-unit tokenization for the Myanmar language (Burmese).

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

Email: ykt.nlp.ai@gmail.com

1.1 Introduction to Sylbreak

မြန်မာစာလုံးတွေကို ဝဏ္ဏအနေနဲ့ဖြတ်နိုင်ဖို့အတွက်က အသုံးများတဲ့ ဝဏ္ဏတွေကို အဘိဓာန်ဆောက်ပြီး ဖြတ်တာမျိုး၊ ဝဏ္ဏတစ်လုံးချင်းစီကို လက်နဲ့သေသေချာချာ ဖြတ်တောက်ထားတဲ့ ကောပတ်စ် ဆောက်ပြီး၊ အဲဒီ ကောပတ်စ်ကိုသုံးပြီး ဆောက်ထားတဲ့ မော်ဒယ် တစ်မျိုးမျိုးနဲ့ ဖြတ်တာမျိုး၊ မြန်မာဝဏ္ဏရဲ့ ဖွဲ့စည်းပုံ ဥပဒေသတွေကိုအခြေခံတဲ့ ဇယားတစ်ခုဆောက်ပြီး အဲဒီဇယားထဲက ဥပဒေသတွေနဲ့ ပရိဂရမ်ရေးဖြတ်တာမျိုး စသည်ဖြင့် အမျိုးမျိုး လုပ်လို့ရပါတယ်။
ကိုယ်တိုင်လည်း အမျိုးမျိုးစမ်းသပ်ပြီး မြန်မာစာကြောင်းတွေကို ဝဏ္ဏဖြတ်တာမျိုးတွေ လုပ်ခဲ့ဖူးပါတယ်။ နောက်ဆုံး အလွယ်ဆုံးဖြစ်ပြီး၊ အသုံးလည်းဝင်တဲ့ Regular Expression (RE) ကို အခြေခံတဲ့ ဖြတ်တဲ့နည်းကိုပဲ သုံးဖြစ်နေခဲ့ပါတယ်။ အများသိအောင် GitHub မှာ ရဲ့မပေးခင် NICT, Kyoto Lab အတွင်းမှာပဲ တစ်ယောက်တည်း သုံးဖြစ်နေခဲ့ပါတယ်။ နာမည်ကိုလည်း sylbreak လို့ပေးထားခဲ့ပါတယ်။ အဲဒီနောက်ပိုင်းမှာ sylbreak4all ဆိုပြီး ယူနီကုဒ်နဲ့ ရိုက်ထားတဲ့တိုင်းရင်းသား ဘာသာစကားတွေကိုပါ ဝဏ္ဏဖြတ်ဖို့ ပရိုပို့ဇယ်တင်ဖြစ်ခဲ့ပါတယ်။ အဲဒါကတော့ ကျွန်တော့်ကျောင်းသားတွေနဲ့ သုတေသန လုပ်ဖော်ကိုင်ဖက်တွေနဲ့ အတူတူ စမ်းသပ်ခဲ့ကြတာပါ။

sylbreak GitHub: https://github.com/ye-kyaw-thu/sylbreak

အောက်ပါအတိုင်း variable တချို့ သတ်မှတ်လိုက်ပြီး

```
my $myConsonant = "က-အ";
my $enChar = "a-zA-Z0-9";
my $otherChar = "ဣဤဥဦဧဩဪ၀-၉၊။!-\/:-\@\[-`{-~\\s";
my $ssSymbol = "္";
my $aThat = "်";
```

RE rule တစ်ကြောင်း ချရေးလိုက်ရင် ဖြတ်လို့ ရပါပြီ။

1.2 Step-by-Step Regex Explanation

Step	Component	Meaning	Example
1	`(?<!$ssSymbol)`	Negative lookbehind for "္"	ရှေ့မှာ ပါဠိဆင့် သင်္ကေတ မရှိ
2	`[$myConsonant]`	Myanmar consonant	မြန်မာဗျည်း စာလုံးများ "က" ကနေ "အ" အထိ
3	`(?![$aThat$ssSymbol])`	Negative lookahead for "်" or "္"	နောက်က အသတ် (သို့) ပါဠိဆင့် သင်္ကေတ
4	`|`	OR	Alternative pattern
5	`[$enChar$otherChar]`	English chars, numbers, symbols	a, 1, !, ကq, etc.

[9]:
```
# import image module
from IPython.display import Image

# get the image
Image(url="./sylbreak_re3.png")
```

[9]:
```
<IPython.core.display.Image object>
```

1.3 sylbreak (perl code)

```
#!/usr/bin/perl

## syllable breaking tool for Myanmar language
## usage: ./sylbreak.pl <-i filename>  [-s separator] [-p {0} or 1]
## e.g. usage1: ./sylbreak.pl -i ../data/my-input
##      usage2: cat ../data/my-input2 | ./sylbreak.pl
##      usage3: ./sylbreak.pl -i ../data/my-input -s "/" -p=1
##
## last updated: 22 July 2016
## added space cleaning parts: 23 May 2022
## Author: Ye Kyaw Thu, Visiting Researcher, Waseda University
## HP:https://sites.google.com/site/yekyawthunlp/

## Reference of Myanmar Unicode: http://unicode.org/charts/PDF/U1000.pdf

use strict;
use warnings;
use utf8;
use Getopt::Long qw(GetOptions);

binmode STDIN, ":encoding(UTF-8)";
binmode STDOUT, ":encoding(UTF-8)";
```

```
my $iOption; #input filename
my $sOption; my $sep = "\|"; #separator
my $pOption; my $printInput = 0; #default value false for print option
my $fh;

GetOptions(
   'help|h|?' => \&help,
   'input-file|i=s' => \$iOption,
   'separator|s=s' => \$sOption,
   'print|p=i' => \$pOption,
) or die "Usage: ./syl-RE-break.pl <-i filename>  [-s separator] [-p {0} or 1]\n";


if ($iOption)
{
  open($fh, '<:encoding(UTF-8)', $iOption) or die "Could not open file '$iOption $!";
}elsif (!defined $iOption )
{
  $fh = *STDIN;
}
else
{
   help();
}


my $myConsonant = "ကာ-အ";
my $enChar = "a-zA-Z0-9";
my $otherChar = "ဣဤဥဦဧဩဪသၦၧ၏၀-၉။!-\/:-\@\[-`{-~\\s";
my $ssSymbol = "္";
my $aThat = "်";


$sep = $sOption if defined ($sOption);
$printInput = $pOption if defined ($pOption);

sub help
{
    print "Usage: ./syl-RE-break.pl <-i filename>  [-s separator] [-p {0} or 1]\n";
    exit(0);
}


while (my $line = <$fh>)
{
   chomp $line;
   if ($printInput == 0)
```

```
    {

        #Regular expression pattern for Myanmar syllable breaking
        #*** a consonant not after a subscript symbol AND a consonant is not followed by a-That
        $line =~ s/((?<!$ssSymbol)[$myConsonant](?![$aThat$ssSymbol])|[$enChar$otherChar])/$sep$
        $line =~ s/^\s+|\s+$//g;
        $line =~ s/ +/ /g;
        print "$line\n";
    }elsif ($printInput == 1)
    {
        print "input: $line\n";
        $line =~ s/((?<!$ssSymbol)[$myConsonant](?![$aThat$ssSymbol])|[$enChar$otherChar])/$sep$
        $line =~ s/^\s+|\s+$//g;
        $line =~ s/ +/ /g;
        print "output: $line\n";
    }

}

close($fh);
```

[10]:
```
!cat ./data/syl/eg.txt
```

စကားပုံဟူသည် အနက်အဓိပ္ပာယ်နှင့် ပြည့်စုံ၍ လူအများ စံပြုပုံခိုင်း
ပြောဆိုလေ့ရှိသော စကား၊ ပြောထုံးစကားဟူ၍ ဖွင့်ဆိုသည်။
(၁) ကြက်ကန်း ဆန်အိုးတိုး
(၂) ကိုင်းကျွန်းမှီ ကျွန်းကိုင်းမှီ
(၃) ကိုယ်ကကျူး ကိုယ့်ဒူးတောင်မယုံရ
(၄) ကိုယ်ထင် ကုတင်ရွှေနန်း
(၅) ကျွဲပါး စောင်းတီး

[13]:
```
!perl ./data/syl/sylbreak.pl -i ./data/syl/eg.txt
```

စ	ကား	ပုံ	ဟူ	သည်		အ	နက်	အ	ဓိပ္ပာယ်	နှင့်		ပြည့်	စုံ	၍		လူ	အ	များ
စံ	ပြု	ပုံ	ခိုင်း		ပြော	ဆို	လေ့	ရှိ	သော		စ	ကား	၊		ပြော	ထုံး	စ	ကား	ဟူ	၍
ဖွင့်	ဆို	သည်	။	
(	၁	)		ကြက်	ကန်း		ဆန်	အိုး	တိုး
(	၂	)		ကိုင်း	ကျွန်း	မှီ		ကျွန်း	ကိုင်း	မှီ
(	၃	)		ကိုယ်	က	ကျူး		ကိုယ့်	ဒူး	တောင်	မ	ယုံ	ရ
(	၄	)		ကိုယ်	ထင်		ကု	တင်	ရွှေ	နန်း
(	၅	)		ကျွဲ	ပါး		စောင်း	တီး

စာလုံးတွေကို ရိုက်တဲ့အခါမှာ Unicode typing order က မှန်ရပါမယ်။
ဥပမာ “ကိုယ့်” ရဲ့ ယ ် ့ အစီအစဉ်ကို မှားယွင်းပြီး ယ ့ ် ဆိုတဲ့ အစီအစဉ်နဲ့ ရိုက်ထည့်ထားပါက
ဝဏ္ဏနှစ်ခုအဖြစ် ဖြတ်ပေးပါလိမ့်မယ်။

[15]:
```
!echo "ကိုယ့်" | perl ./data/syl/sylbreak.pl
```

| ကို | ယ့်

1.4 oppaWord

Coding အပိုင်း oppaWord ရဲ့ algorithm အပိုင်းကို ရှင်းမယ်ဆိုရင် အချိန်ထပ်ယူရပါလိမ့်မယ်။
ဒီနေရာမှာတော့ လက်တွေ့ လုပ်တဲ့ အပိုင်းကိုပဲ ဦးစားပေးကြည့်ရအောင်။

အသေးစိတ်က အောက်ပါလင့် ကနေ လေ့လာပါလို့ အကြံပေးချင်ပါတယ်။
GitHub Link: https://github.com/ye-kyaw-thu/oppaWord

oppaWord အလုပ်လုပ်ပုံကို visualization လုပ်ပြရရင်တော့ အောက်ပါအတိုင်းပါ။

[17]:
```
# import image module
from IPython.display import Image

# get the image
Image(url="./overview-of-oppaWord.png")
```

[17]:
```
<IPython.core.display.Image object>
```

[16]:
```
! python /home/ye/exp/myTokenizer/oppaWord/oppa_word.py --help
```

```
usage: oppa_word.py [-h] --input INPUT [--output OUTPUT] --dict DICT
                    [--sylfreq SYLFREQ] [--arpa ARPA]
                    [--postrule-file POSTRULE_FILE] [--max-order MAX_ORDER]
                    [--dict-weight DICT_WEIGHT] [--use-bimm-fallback]
                    [--bimm-boost BIMM_BOOST] [--visualize-dag]
                    [--dag-output-dir DAG_OUTPUT_DIR]
                    [--space-remove-mode {all,my,my_not_num}]
                    [--max-word-len MAX_WORD_LEN]

oppa_word, Hybrid DAG + BiMM + LM Myanmar Word Segmenter with optional Aho-
Corasick support

options:
  -h, --help            show this help message and exit
  --input INPUT, -i INPUT
                        Input file with one sentence per line (UTF-8)
  --output OUTPUT, -o OUTPUT
                        Optional output file path (default: stdout)
  --dict DICT, -d DICT  Word dictionary file (one word per line)
  --sylfreq SYLFREQ, -s SYLFREQ
                        Syllable frequency file (syllable<TAB>frequency, for
                        scoring)
  --arpa ARPA, -a ARPA  ARPA-format syllable-level language model (optional)
  --postrule-file POSTRULE_FILE
                        Optional post-processing rules (e.g., merging,
```

```
                        corrections)
  --max-order MAX_ORDER
                        Max LM n-gram order (default: 5)
  --dict-weight DICT_WEIGHT
                        Dictionary path weight in scoring (default: 10.0)
  --use-bimm-fallback   Enable Bi-directional Maximum Matching as fallback
  --bimm-boost BIMM_BOOST
                        Boost score added to Bi-MM fallback path (default:
                        0.0)
  --visualize-dag       Generate DAG visualization (PDF per sentence)
  --dag-output-dir DAG_OUTPUT_DIR
                        Directory to save DAG PDFs if --visualize-dag is used
                        (default: 'dag_viz')
  --space-remove-mode {all,my,my_not_num}
                        Preprocessing mode to remove spaces: 'all', 'my'
                        (Myanmar only), or 'my_not_num (Myanmar but not
                        including Myanmar numbers'
  --max-word-len MAX_WORD_LEN
                        Maximum word length in syllables (3-12, default:6)
```

[19]:
```
! time python /home/ye/exp/myTokenizer/oppaWord/oppa_word.py \
  --input ./data/syl/eg.txt \
  --dict /home/ye/exp/myTokenizer/oppaWord/data/myg2p_mypos.dict \
  --arpa /home/ye/exp/myTokenizer/oppaWord/data/myMono_clean_syl.trie.bin \
  --use-bimm-fallback \
  --bimm-boost 150 \
  --space-remove-mode "my_not_num"
```

စကားပုံ ဟူသည် အနက်အဓိပ္ပာယ် နှင့် ပြည့် စုံ၍ လူ အများ စံ ပြု ပုံခိုင်း ပြောဆို လေ့ ရှိ သော စကား ၊ ပြော ထုံး စကား ဟူ၍ ဖွင့်ဆို သည် ။
(၁) ကြက် ကန်း ဆန် အိုး တိုး
(၂) ကိုင်းကျွန်းမှီကျွန်းကိုင်းမှီ
(၃) ကိုယ် က ကျူး ကိုယ့် ဒူး တောင် မ ယုံ ရ
(၄) ကိုယ် ထင် ကုတင် ရွေ နန်း
(၅) ကျွဲ ပါး စောင်း တီး

```
real    0m0.066s
user    0m0.044s
sys     0m0.022s
```

[23]:
```
!cat ./data/mgyin.txt
```

ထိုအခါ မောင်ရင်မောင်စိတ်တွင်း၌ စဉ်းစားဆင် ခြင်သည်ကား ယခုအခါ ရည်းစားဖြစ်သူ မမယ်မကို ဦးစွာနှုတ်ဆက်ရမည် သို့တည်းမဟုတ် အမိဖြစ်သူကို ဦးစွာနှုတ်ဆက်ရမည်မှာ ငါမပိုင်းမဖြတ်နိုင်အောင် ရှိတော့သည်။ မမယ်မလည်း ငါ သူကို ဦးစွာနှုတ်ခွန်းဆက်မည်၊ အမိအား ဦးစွာနှုတ်ခွန်းဆက်မည်ကို သတိပြုလျက် နေကောင်းနေလိမ့်မည်။ သို့သော်လည်း အမိဖြစ်သူအဖို့ မကောင်းသတင်းပါသည်။ အမိကိုသာလျှင် ဦးစွာနှုတ်ဆက်အံ့ဟု ကြံပြီးလျှင်

မောင်ရင်မောင်က ဆိုသည်မှာ အမေ့-အဖေကား မပါလာပြီ၊ မြင်းခြံ အောက်တွင်
ကျန်ရစ်ခဲ့လေပြီဆို၏။ ထိုအခါ မဖားဥ ကလည်း အလို-မောင်ရင်မောင်၊ ဘယ်အကြောင်း ကြောင့်
မင့်အဖ ကျန်ရစ်ခဲ့ပါသနည်း၊ အမိအား ကုန်စင်အောင်ပြောပါဆို၏။ မောင်ရင်မောင်လည်း
အမေ-ယခုအကြောင်းကား အလွန်ထူးခြားသော အကြောင်းဖြစ်သည်၊ ကျွန်တော့်အဖမှာ မလွဲသာသော
ကြောင့် နေရစ်ခဲ့ရှာရသည် မောင်ရင်မောင်ကဆိုလျှင် မောင်ရင်မောင် ဤအမှုကား
ထူးခြားလှသည်၊ မောင့်အဖသည် မည်သည့်နေ့ရက်လိုက်ပါလာခဲ့မည် မှာလိုက်သနည်း၊
ယခုလပြည့်ကျော်တစ်ရက်၊ နှစ်ရက် အတွင်း ရောက်မည်လော၊ မရောက်မည်လော၊ လာမည်လော၊
မလာမည်လော၊ ကိုယ့်လှေပိုင်ရှိလျက်၊ ကိုယ့်သားပဲ့နင်းမှန်လျက်၊
ကိုယ့်သားပဲ့နင်းလုပ်သည့် လှေကို မစီးမနင်းလို နေရစ်ခဲ့သည်ကား လွန်စွာထူးခြားသည်၊
အထူးအရေးကြီးသောကိစ္စ တစ်စုံတစ်ရာရှိအံ့၊ ငါ့သား မောင်ရင်မောင်သိရှိသမျှသောကိစ္စကို
အမေအားပြောကြားပါ ဆိုလျှင် မောင်ရင်မောင်လည်း မမှန်သည့်စကားကို မဆိုသင့်ပြီ၊
အမှန်ကိုပြောမှ သင့်မည်ကြံပြီးလျှင် ဤသို့ပြော၏။ အမေ-အမှန်ကို ကျွန်တော်
ပြောပြရမည်ဆိုသော် ကျွန်တော်တို့အဖသည် မြင်းခြံမြို့အောက် မြင်ကွန်းမြို့၌
ကျွန်တော်တို့လှေဆိုက်ကပ်နေသည့်အတွင်း လူအပေါင်းတို့တွင် ကျရောက် လျက်ရှိသော
ကပ်ရောဂါကြီး ဖိစီးနှိပ်စက်သည့်အတွက် ကြောင့် အသက်ဆုံးရှုံးလေပြီ ဟူ၍ဆိုသည်တွင်
ထိုမိန်းမအို မဖားဥသည် ရုတ်တရက် နောက်သို့ပြန်၍ လဲကျသောအခါ မောင်ရင်မောင်
ပြေးသွားဖက်မည် အလုပ်တွင် မမယ်မက ဝိုင်ကြွေးလျက် ကောက်ယူပွေ့ပိုက်ပြုစုသောကြောင့်
ဤနေရာကား မိန်းမနေရာ ဖြစ်သည်။ ငါဝင်ရောက်ပြုစုရန်မတော်ဟု မောင်ရင်မောင်သည်
ရပ်တန့်လျက်နေလေ၏။

[24]:
```
! time python /home/ye/exp/myTokenizer/oppaWord/oppa_word.py \
  --input ./data/mgyin.txt \
  --dict /home/ye/exp/myTokenizer/oppaWord/data/myg2p_mypos.dict \
  --arpa /home/ye/exp/myTokenizer/oppaWord/data/myMono_clean_syl.trie.bin \
  --use-bimm-fallback \
  --bimm-boost 150 \
  --space-remove-mode "my_not_num"
```

ထိုအခါ မောင်ရင် မောင် စိတ် တွင်း၌ စဉ်းစား ဆင်ခြင် သည် ကား ယခု အခါ ရည်းစား ဖြစ်
သူမ မယ် မ ကိုဦး စွာ နှုတ်ဆက် ရ မည် သို့တည်းမဟုတ် အမိ ဖြစ် သူ ကိုဦး စွာ နှုတ်ဆက်
ရ မည် မှာ ငါ မ ပိုင်း မ ဖြတ် နိုင် အောင် ရှိ တော့ သည် ။ မ မယ် မ လည်း ငါ သူ
ကိုဦး စွာ နှုတ်ခွန်းဆက် မည် ၊ အမိ အားဦး စွာ နှုတ်ခွန်းဆက် မည် ကို သတိပြု လျက်
နေကောင်း နေ လိမ့်မည် ။ သို့သော်လည်း အမိ ဖြစ် သူ အဖို့ မကောင်းသတင်း ပါ သည် ။ အမိ
ကို သာ လျှင်ဦး စွာ နှုတ်ဆက် အံ့ ဟု ကြံ ပြီးလျှင် မောင်ရင် မောင် က ဆို သည်မှာ အ
မေ့- အဖေ ကား မ ပါ လာ ပြီ ၊ မြင်းခြံ အောက် တွင် ကျန်ရစ် ခဲ့ လေ ပြီ ဆို၏ ။ ထိုအခါ
မ ဖားဥ က လည်း အ လို- မောင်ရင် မောင် ၊ ဘယ် အကြောင်း ကြောင့် မင့် အဖ ကျန်ရစ် ခဲ့
ပါ သနည်း ၊ အမိ အား ကုန်စင်အောင် ပြော ပါ ဆို၏ ။ မောင်ရင် မောင် လည်း အ မေ- ယခု
အကြောင်း ကား အလွန် ထူးခြား သော အကြောင်း ဖြစ် သည် ၊ ကျွန်တော့် အဖ မှာ မ လွဲ သာ
သောကြောင့် နေရစ် ခဲ့ ရှာ ရ သည် မောင်ရင် မောင် က ဆိုလျှင် မောင်ရင် မောင်ဤ အမှု
ကား ထူးခြား လှ သည် ၊ မော င့် အဖ သည် မည် သည့် နေ့ရက် လိုက်ပါ လာ ခဲ့ မည် မှာ ␣
↪လိုက်
သနည်း ၊ ယခု လပြည့် ကျော် တစ် ရက် ၊ နှစ် ရက် အတွင်း ရောက် မည် လော ၊ မ ရောက်␣
↪မည်

လော ၊ လာ မည် လော ၊ မလာ မည် လော ၊ ကိုယ့် လေ့ ပိုင် ရှိ လျက် ၊ ကိုယ့် သား ပဲ့နင်း မှန် လျက် ၊ ကိုယ့် သား ပဲ့နင်း လုပ် သည့် လေ့ ကို မ စီး မ နင်း လို နေရစ် ခဲ့ သည် ကား လွန်စွာ ထူးခြား သည် ၊ အထူး အရေးကြီးသော ကိစ္စ တစ်စုံတစ်ရာ ရှိ အံ့ ၊ ငါ့ သား မောင်ရင် မောင် သိရှိ သမျှ သော ကိစ္စ ကို အမေ အား ပြောကြား ပါ ဆိုလျှင် မောင်ရင် မောင် လည်း မ မှန် သည့် စကား ကို မဆို သင့် ပြီ ၊ အမှန် ကို ပြော မှ သင့် မည် ကြံ ပြီး လျှင်ဤ သို့ ပြော၏ ။ အ မေ- အမှန် ကို ကျွန်တော် ပြောပြ ရ မည် ဆိုသော် ကျွန်တော် တို့ အဖ သည် မြင်းခြံ မြို့ အောက် မြင် ကွန်း မြို့၌ ကျွန်တော် တို့ လေ့ ဆိုက်ကပ် နေ သည့် အတွင်းလူ အပေါင်း တို့ တွင် ကျရောက် လျက် ရှိ သော ကပ် ရောဂါကြီး ဖိစီး နှိပ်စက် သည့်အတွက် ကြောင့် အသက် ဆုံးရှုံး လေ ပြီ ဟူ၍ ဆို သည် တွင် ထို မိန်းမ အို မ ဖားဥ သည် ရုတ်တရက် နောက် သို့ ပြန်၍ လဲကျ သောအခါ မောင်ရင် မောင် ပြေး သွား ဖက် မည် အလုပ်တွင် မ မယ် မက ငိုကြွေး လျက် ကောက်ယူ ပွေ့ပိုက် ပြုစု သော ကြောင့်ဤ နေရာ ကား မိန်းမ နေရာ ဖြစ် သည် ။ ငါ ဝင်ရောက် ပြုစု ရန် မတော် ဟု မောင်ရင် မောင် သည် ရုပ်တန့် လျက် နေ လေ၏ ။

```
real    0m0.071s
user    0m0.041s
sys     0m0.030s
```

1.5 Notes

- NLP အလုပ်တွေအပေါ်ကို မူတည်ပြီး တချို့အလုပ်တွေအတွက် ဝဏ္ဏကိုဖြတ်ပြီး လုပ်ကြတယ်။ တချို့သော အလုပ်တွေက စာလုံး (word level) သေချာဖြတ်ထားမှ သင့်တော်တဲ့ အပိုင်း ရှိပါတယ်။
- LLM တို့လို pretrained model တွေကို သုံးတဲ့အခါမှာတော့ အဲဒီမော်ဒယ်ကို ဆောက်စဉ်က သုံးထားတဲ့ ယူနစ် (အများအားဖြင့် subword ဖြစ်ပြီး ဘာသာစကားအများကြီးအတွက် အဆင်ပြေမယ့် ပုံစံနဲ့ဖြတ်ထားတာမျိုးလည်း ရှိပါတယ်)
- sylbreak ရော oppaword နှစ်မျိုးစလုံးကို အသုံးပြုကြည့်ကြပါ
- အထူးသဖြင့် oppaword က စာလုံးဖြတ်မှားတာမျိုးကို post-editing RE နဲ့ ပြန်ပြင်ခိုင်းလို့ ရပါတယ်။ ပြီးတော့ စာလုံးဖြတ်ပေးတဲ့ နှုန်းက တအားမြန်တာမို့ အသုံးဝင်ပါလိမ့်မယ်
- လက်ရှိထက် စာလုံးဖြတ်တာ မှန်ကန်ဖို့အတွက်က လူက လက်နဲ့သေသေချာချာ မှန်မှန်ကန်ကန် စာလုံးဖြတ်ထားဖို့ လိုအပ်ပါတယ်
- ပြီးတော့ မြန်မာစာအတွက် ဘယ်လိုပုံစံနဲ့ တသတ်မှတ်တည်း ဖြစ်ရန်ဆိုတဲ့ ဥပဒေသလည်း သတ်မှတ်ပေးနိုင်ရင် ပိုကောင်းပါလိမ့်မယ်

[ ]: