exp_with_Dictionary

August 3, 2025

0.1 Experriment with Bigger Dictionary

ဒီတခေါက် မှာတော့ myG2P+myPOS အဘိဓာန်ကို myRoman က နာမည်တွေအပြင် ထပ်ဖြည့်ထားတဲ့ နာမည်တွေနဲ့ ဆောက်ထားတဲ့ name dictionary ကိုပါ ပေါင်းလိုက်ပြီး ပိုကြီးတဲ့ dictionary ကိုသုံးပြီး word segmentation experiment လုပ်ကြည့်မှာ ဖြစ်ပါတယ်။

Date: 3 Aug 2025
Run by Ye Kyaw Thu, LU Lab., Myanmar

[1]:
```
!pwd
```

```
/home/ye/exp/myTokenizer
```

[2]:
```
cd /home/ye/exp/myTokenizer/oppaWord/data/prepare_big_dict/combine
```

```
/home/ye/exp/myTokenizer/oppaWord/data/prepare_big_dict/combine
```

[3]:
```
!ls
```

```
myg2p_mypos_name.dict
```

[5]:
```
!mv ./myg2p_mypos_name.dict ./myg2p_mypos_name.txt
```

[6]:
```
!wc myg2p_mypos_name.txt
```

```
 155538  155538 5261962 myg2p_mypos_name.txt
```

[7]:
```
!sort ./myg2p_mypos_name.txt | uniq > ./myg2p_mypos_name.dict
```

[8]:
```
!wc ./myg2p_mypos_name.dict
```

```
 154684  154684 5244464 ./myg2p_mypos_name.dict
```

[11]:
```
!cp ./data/prepare_big_dict/combine/myg2p_mypos_name.dict ./data/
```

0.2 oppaWord with myg2p_mypos.dict (Open-test)

[10]:
```
cd /home/ye/exp/myTokenizer/oppaWord/
```

```
/home/ye/exp/myTokenizer/oppaWord
```

[14]:
```
!mkdir exp_2
```

[15]:
```
!time python oppa_word.py \
  --input "./data/otest.1k.word" \
  --output "exp_2/dict_only_bimmfallback_bimmboost150_otest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos.dict" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.115s
user    0m0.103s
sys     0m0.012s
```

[16]:
```
cd exp_2
```

```
/home/ye/exp/myTokenizer/oppaWord/exp_2
```

[17]:
```
!wc ./dict_only_bimmfallback_bimmboost150_otest.seg
```

```
  1000  12856 180176 ./dict_only_bimmfallback_bimmboost150_otest.seg
```

[18]:
```
!head ./dict_only_bimmfallback_bimmboost150_otest.seg
```

```
တစ် ကိုက် ကို ဝမ် ခုနှစ်ထောင် ပါ ။
မနှစ် က သူ ကျွန်မ ကို သင်ပေး တယ် ။
ကျွန်တော့် ခုံ သွား ရှာ မလို့ ။
အတန်း စတာ ကြာ ပြီ လား ။
ဆေး နည်းနည်း စား လိုက် ၊ သုံး လေး ရက် လောက် အနားယူ လိုက် ရင် ပျောက် သွား မှာ ပါ
။
အေးချမ်း မှု နဲ့ စည်းကမ်း ကို တည်မြဲ အောင် ထိန်းသိမ်း သည် ။
ဇွန်း ကို လိုအပ် တယ် ။
ဘွဲ့ရ ရင် ဘာ လုပ် မလို့ လဲ ။
ကျွန်တော် ချောင်းဆိုး ခြင်း အတွက် တစ်ခုခု လို ချင်တယ် ။
အသီးအနှံ တို့ မှ လွဲလျှင် လူ တို့၏ အဓိက အစားအစာ မှာ ငါး ဖြစ် သည် ။
```

[19]:
```
!tail ./dict_only_bimmfallback_bimmboost150_otest.seg
```

```
အိုးခွက်ပန်းကန် တွေ သိပ် မရှိ လို့ ထမင်းဟင်း ချက် ရ တာ အဆင်မပြေ ဘူး ။
စိတ်ဝင်စား ဖို့ ကောင်း တယ် ။
ဒီ ဆေး ကို တဝက် စီ ခွဲ ပေး ပါ ။
ရောင်း ကောင်း လား ။
ဆရာ ဒီ သွား က ခဏခဏ နာ နေ တယ် ။
အခု ဘာ လုပ် နေ လဲ ။
ဇူ လိုင် ၁၄ ရက် မှာ ဘန်ကောက် ကို သွား မယ့် US 123 မှာ ပါ ။ ဟုတ်လား ။
ကား မှ ကားဘီး ကို ဖြုတ် လိုက် သည် ။
```

ကျွန်တော် သိ ပါရစေ ။
ဘူတာရုံ က အလွန်တရာ ပြည့်ကျပ် နေ သည် ။

0.3 Evaluation

[20]:
```
cd ..
```

```
/home/ye/exp/myTokenizer/oppaWord
```

[22]:
```
!ls ./tools/
```

```
correct_my_punc.py  eval_segmentation.py  evaluate.py
```

[23]:
```
!python ./tools/eval_segmentation.py -r ./data/otest.1k.word -H ./exp_2/
 ↪dict_only_bimmfallback_bimmboost150_otest.seg --top-k 10
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.8857
Word Recall                    0.8454
Word F1-score                  0.8651
------------------------------------------------------------
Boundary Precision             0.5516
Boundary Recall                0.5265
Boundary F1-score              0.5387
------------------------------------------------------------
Vocab Precision                0.8139
Vocab Recall                   0.9184
Vocab F1-score                 0.8630
============================================================

Additional Statistics:
Reference words: 13468
Hypothesis words: 12856
Correct words: 11386
Reference vocabulary size: 2709
Hypothesis vocabulary size: 3057
Common vocabulary: 2488

Top Segmentation Errors Analysis
============================================================
Total errors: 5665

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):
```

```
Most Frequent Complex Boundary Errors:
   81 × REF: 'တယ်' → HYP: '။'
   74 × REF: 'ပါ' → HYP: '။'
   57 × REF: 'သည်' → HYP: '။'
   42 × REF: 'ပါ' → HYP: 'ပါတယ်'
   24 × REF: 'မယ်' → HYP: '။'
   18 × REF: 'ဘူး' → HYP: '။'
   18 × REF: 'လဲ' → HYP: '။'
   17 × REF: 'လား' → HYP: '။'
   16 × REF: 'ဖြစ်' → HYP: 'သည်'
   15 × REF: 'ခဲ့' → HYP: 'သည်'
```

0.4 oppaWord with myg2p_mypos.dict + post_rules

[26]:
```
!wc ./data/rules.txt
```

```
  31   62 1250 ./data/rules.txt
```

[27]:
```
!cat ./data/rules.txt
```

```
ပါတယ်|||ပါ တယ်
မရှိ|||မ ရှိ
ဒီနေ့|||ဒီ နေ့
ဖြစ်သည်|||ဖြစ် သည်
ခဲ့သည်|||ခဲ့ သည်
သူက|||သူ က
မဟုတ်|||မ ဟုတ်
များကို|||များ ကို
ပါဘူး|||ပါ ဘူး
ကတော့|||က တော့
ရှိတယ်|||ရှိ တယ်
ချင်ပါ|||ချင် ပါ
ချင်တယ်|||ချင် တယ်
မြန်မာနိုင်ငံ|||မြန်မာ နိုင်ငံ
တို့၏|||တို့ ၏
ပါသလဲ|||ပါ သလဲ
မှာရှိ|||မှာ ရှိ
များ၏|||များ ၏
ထင်တယ်|||ထင် တယ်
မှာထား|||မှာ ထား
စီးပွားရေး|||စီးပွား ရေး
အလုပ်လုပ်|||အလုပ် လုပ်
တာနဲ့|||တာ နဲ့
```

လိုက်ပါ|||လိုက် ပါ
ဒီဟာ|||ဒီ ဟာ
သူများ|||သူ များ
သူ၏|||သူ ၏
ဟုတ်လား|||ဟုတ် လား
တစ်ဦး|||တစ် ဦး
နံ ပါတ်|||နံပါတ်
(\S)([။၊])|||\1 \2

[28]:
```
!time python oppa_word.py \
  --input "./data/otest.1k.word" \
  --output "exp_2/dict_rules_bimmfallback_bimmboost150_otest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos.dict" \
  --postrule-file "./data/rules.txt" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.108s
user    0m0.099s
sys     0m0.009s
```

[29]:
```
!wc ./exp_2/dict_rules_bimmfallback_bimmboost150_otest.seg
```

```
  1000  13170 180502 ./exp_2/dict_rules_bimmfallback_bimmboost150_otest.seg
```

[30]:
```
!head ./exp_2/dict_rules_bimmfallback_bimmboost150_otest.seg
```

တစ် ကိုက် ကို ဝမ် ခုနှစ်ထောင် ပါ ။
မနှစ် က သူ ကျွန်မ ကို သင်ပေး တယ် ။
ကျွန်တော့် ခုံ သွား ရှာ မလို့ ။
အတန်း စတာ ကြာ ပြီ လား ။
ဆေး နည်းနည်း စား လိုက် ၊ သုံး လေး ရက် လောက် အနားယူ လိုက် ရင် ပျောက် သွား မှာ ပါ ။
အေးချမ်း မှု နဲ့ စည်းကမ်း ကို တည်မြဲ အောင် ထိန်းသိမ်း သည် ။
ဖွန်း ကို လိုအပ် တယ် ။
ဘွဲ့ရ ရင် ဘာ လုပ် မလို့ လဲ ။
ကျွန်တော် ချောင်းဆိုး ခြင်း အတွက် တစ်ခုခု လို ချင် တယ် ။
အသီးအနှံ တို့ မှ လွဲလျှင် လူ တို့ ၏ အဓိက အစားအစာ မှာ ငါး ဖြစ် သည် ။

[31]:
```
!tail ./exp_2/dict_rules_bimmfallback_bimmboost150_otest.seg
```

အိုးခွက်ပန်းကန် တွေ သိပ် မ ရှိ လို့ ထမင်းဟင်း ချက် ရ တာ အဆင်မပြေ ဘူး ။
စိတ်ဝင်စား ဖို့ ကောင်း တယ် ။
ဒီ ဆေး ကို တဝက် စီ ခွဲ ပေး ပါ ။
ရောင်း ကောင်း လား ။

ဆရာ ဒီ သွား က ခဏခဏ နာ နေ တယ် ။
အခု ဘာ လုပ် နေ လဲ ။
ဇူ လိုင် ၁၄ ရက် မှာ ဘန်ကောက် ကို သွား မယ့် US 123 မှာ ပါ ။ ဟုတ် လား ။
ကား မှ ကားဘီး ကို ဖြုတ် လိုက် သည် ။
ကျွန်တော် သိ ပါရစေ ။
ဘူတာရုံ က အလွန်တရာ ပြည့်ကျပ် နေ သည် ။

0.5 Evaluation

[32]:
```
!python ./tools/eval_segmentation.py -r ./data/otest.1k.word -H ./exp_2/
↪dict_rules_bimmfallback_bimmboost150_otest.seg --top-k 10
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.9094
Word Recall                    0.8893
Word F1-score                  0.8992
------------------------------------------------------------
Boundary Precision             0.6116
Boundary Recall                0.5981
Boundary F1-score              0.6048
------------------------------------------------------------
Vocab Precision                0.8183
Vocab Recall                   0.9158
Vocab F1-score                 0.8643
============================================================

Additional Statistics:
Reference words: 13468
Hypothesis words: 13170
Correct words: 11977
Reference vocabulary size: 2709
Hypothesis vocabulary size: 3032
Common vocabulary: 2481

Top Segmentation Errors Analysis
============================================================
Total errors: 4966

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):

Most Frequent Complex Boundary Errors:
```

```
50 × REF: 'တယ်' → HYP: '။'
49 × REF: 'သည်' → HYP: '။'
39 × REF: 'ပါ' → HYP: '။'
20 × REF: 'ပါ' → HYP: 'တယ်'
19 × REF: '။' → HYP: 'သည်'
19 × REF: 'မယ်' → HYP: '။'
17 × REF: 'လား' → HYP: '။'
17 × REF: 'ဖြစ်' → HYP: 'သည်'
13 × REF: 'ဘူး' → HYP: '။'
12 × REF: 'လဲ' → HYP: '။'
```

0.6 oppaWord with myg2p_mypos.dict (Closed-test)

[33]:
```
!time python oppa_word.py \
  --input "./data/10k_test.txt" \
  --output "exp_2/dict_only_bimmfallback_bimmboost150_ctest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos.dict" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.628s
user    0m0.614s
sys     0m0.014s
```

[34]:
```
!wc ./exp_2/dict_only_bimmfallback_bimmboost150_ctest.seg
```

```
  10000  112271 1680793 ./exp_2/dict_only_bimmfallback_bimmboost150_ctest.seg
```

[35]:
```
!head ./exp_2/dict_only_bimmfallback_bimmboost150_ctest.seg
```

```
၁၉၆၂ ခုနှစ် ခန့်မှန်း သန်းခေါင်စာရင်း အရ လူဦး ရေ ၁၁၅၉၃၁ ယောက် ရှိ သည်
လူ တိုင်း တွင် သင့်မြတ် လျော်ကန် စွာ ကန့်သတ် ထား သည့် အလုပ်လုပ်ချိန် အပြင် လစာ
နှင့်တကွ အခါ ကာလ အားလျော်စွာ သတ်မှတ် ထား သည့် အလုပ် အားလပ်ရက် များ ပါဝင် သည့်
အနားယူခွင့် နှင့် အားလပ်ခွင့် ခံစားပိုင်ခွင့် ရှိ သည်
ဤ နည်း ကို စစ်ယူ သော နည်း ဟု ခေါ် သည်
စာပြန်ပွဲ ဆို တာ က အာဂုံဆောင် အလွတ်ကျက် ထား တဲ့ ပိဋကတ်သုံးပုံ စာပေ တွေ ကို စာစစ်
သံဃာတော်ကြီး တွေ ရဲ့ ရှေ့မှာ အလွတ် ပြန် ပြီး ရွတ်ပြ ရ တာ ပေါ့
ဒီ မှာ ကျွန်တော့် သက်သေခံကတ် ပါ
၂ ဝ ရာစု မြန်မာ့ သမိုင်း သန်း ဝင်း လှိုင် ၂ ဝ ဝ၉ ခု မေ လ ကံကော်ဝတ်ရည် စာပေ
ကျွန်တော် မျက်မှန် တစ် လက်လုပ် ချင်ပါတယ်
ကျွန်တော် တို့ က ဒီ အမှု ရဲ့ ကြံရာပါ ကို ဖမ်းမိ ဖို့ ကြိုးစား ခဲ့ တယ်
ကလေး မီးဖွား ဖို့ ခန့်မှန်း ရက် က ဘယ်တော့ ပါ လဲ
အရိုးရှင်းဆုံး ကာဗိုဟိုက်ဒရိတ် မှာ ဂလူးကို့စ် ဂလက်တို့စ် ဖရုပ်တို့စ် စသည့်
မိုနိုဆက်ကရိုက် များ ဖြစ် သည်
```

[36]:
```
!tail ./exp_2/dict_only_bimmfallback_bimmboost150_ctest.seg
```

ကျွန်တော် စိန် နဲ့ တစ် ခုလို ချင်ပါတယ်
ကြိုးကြာ
သူ အဘိဓာန်စာလုံး ရှာ တတ် ပါတယ်
သြော် ကျွန်တော် မျက်စိလည် နေ ပြီ ထင်တယ် ဒီ နေရာ က နေ ဘူတာရုံ ကို ဘယ်လို သွား ရ␣
↪မလဲ
လော့အိန်ဂျလိစ် က နေ တိုကျို ကို လေယာဉ်ခ ဘယ်လောက် လဲ
ပထမဆုံး အဆင်ပြေ မယ့် လေယာဉ် လို ချင်ပါတယ်
ရေထွက်ပစ္စည်း များ ထုတ်လုပ် မှု တွင် လည်း တန် ချိန် ၄၁.၂၄၄ သန်း ထိ ၁၉၇၈ ခုနှစ်
ထက် ၈.၈ ဆတိုး တက် ထုတ်လုပ် နိုင် ခဲ့ သည်
ဟုတ်တယ် ဟင်းသီးဟင်းရွက် သွား ဝယ် ရအောင်
ဆရာကြီး က ၄ နာရီ လောက် ဆို အလုပ် နည်းနည်း ရှင်း ပြီ ၅ နာရီ မှ ရုံးဆင်း မှာ ဆို
တော့ တစ်နာရီ လောက် တော့ တွေ့ ချိန် ရ မှာ ပဲ
ဝင်ခွင့် ဈေးနှုန်း ရော ပါ လား

0.7 Check Codepoints

[42]:
```
!perl ./tools/print-codepoint.pl ./exp_2/chk_wa_zero.txt
```

၂ ဝ ရာစု မြန်မာ့ သမိုင်း သန်း ဝင်း လှိုင် ၂ ဝ ဝ၉ ခု မေ လ ကံကော်ဝတ်ရည် စာပေ
၂ (4162, U1042) (32, U20) ဝ (4125, U101d) (32, U20) ရ (4123, U101b) ◌ာ␣
↪(4140,
U102c) စ (4101, U1005) ◌ု (4143, U102f) (32, U20) မ (4121, U1019) ◌ြ (4156,
U103c) န (4116, U1014) ◌် (4154, U103a) မ (4121, U1019) ◌ာ (4140, U102c) ◌့␣
↪(4151,
U1037) (32, U20) သ (4126, U101e) မ (4121, U1019) ◌ိ (4141, U102d) ◌ု (4143,
U102f) ဂ (4100, U1004) ◌် (4154, U103a) ◌း (4152, U1038) (32, U20) သ (4126,
U101e) န (4116, U1014) ◌် (4154, U103a) ◌း (4152, U1038) (32, U20) ဝ (4125,
U101d) ဂ (4100, U1004) ◌် (4154, U103a) ◌း (4152, U1038) (32, U20) လ (4124,
U101c) ◌ှ (4158, U103e) ◌ိ (4141, U102d) ◌ု (4143, U102f) ဂ (4100, U1004) ◌်␣
↪(4154,
U103a) (32, U20) ၂ (4162, U1042) (32, U20) ဝ (4125, U101d) (32, U20) ဝ
(4125, U101d) ၉ (4169, U1049) (32, U20) ခ (4097, U1001) ◌ု (4143, U102f) ␣
↪(32,
U20) မ (4121, U1019) ◌ေ (4145, U1031) (32, U20) လ (4124, U101c) (32, U20) ␣
↪က
(4096, U1000) ◌ံ (4150, U1036) က (4096, U1000) ◌ေ (4145, U1031) ◌ာ (4140,␣
↪U102c) ◌်
(4154, U103a) ဝ (4125, U101d) တ (4112, U1010) ◌် (4154, U103a) ရ (4123, U101b)␣
↪ည
(4106, U100a) ◌် (4154, U103a) (32, U20) စ (4101, U1005) ◌ာ (4140, U102c) ပ
(4117, U1015) ◌ေ (4145, U1031) , no. of char = 74
၀၀၇

ం (4160, U1040) ం (4160, U1040) ా (4167, U1047) , no. of char = 3

[45]:
```
!python ./tools/eval_segmentation.py --help
```

```
usage: eval_segmentation.py [-h] -r REFERENCE [-H HYPOTHESIS] [--top-k TOP_K]
                            [--no-errors]

Enhanced Word Segmentation Evaluator with Error Analysis

options:
  -h, --help            show this help message and exit
  -r REFERENCE, --reference REFERENCE
                        Reference (gold standard) file (default: None)
  -H HYPOTHESIS, --hypothesis HYPOTHESIS
                        Hypothesis (system output) file (use - for stdin)
                        (default: -)
  --top-k TOP_K         Show top K most frequent errors (default: 10)
  --no-errors           Skip error analysis to save time (default: False)
```

[46]:
```
!python ./tools/eval_segmentation.py -r ./data/10k_test.txt -H ./exp_2/
↪dict_only_bimmfallback_bimmboost150_ctest.seg
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.8795
Word Recall                    0.8378
Word F1-score                  0.8582
------------------------------------------------------------
Boundary Precision             0.5635
Boundary Recall                0.5368
Boundary F1-score              0.5498
------------------------------------------------------------
Vocab Precision                0.8242
Vocab Recall                   0.9228
Vocab F1-score                 0.8707
============================================================

Additional Statistics:
Reference words: 117857
Hypothesis words: 112271
Correct words: 98746
Reference vocabulary size: 10840
Hypothesis vocabulary size: 12137
Common vocabulary: 10003

Top Segmentation Errors Analysis
```

```
============================================================
Total errors: 47862

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):

Most Frequent Complex Boundary Errors:
  430 × REF: 'ပါ' → HYP: 'ပါတယ်'
  119 × REF: 'ဖြစ်' → HYP: 'သည်'
  118 × REF: 'ဒီ' → HYP: 'ဒီနေ့'
  108 × REF: 'မ' → HYP: 'မရှိ'
   91 × REF: 'သူ' → HYP: 'သူက'
   81 × REF: 'ခဲ့' → HYP: 'သည်'
   81 × REF: 'မ' → HYP: 'မဟုတ်'
   71 × REF: 'က' → HYP: 'ကတော့'
   68 × REF: 'ရှိ' → HYP: 'ရှိတယ်'
   66 × REF: 'ခဲ့' → HYP: 'တယ်'
```

0.8 oppaWord with myg2p_mypos.dict+rules (Closed-test)

[47]:
```
!time python oppa_word.py \
  --input "./data/10k_test.txt" \
  --output "exp_2/dict_rules_bimmfallback_bimmboost150_ctest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos.dict" \
  --postrule-file "./data/rules.txt" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.662s
user    0m0.650s
sys     0m0.012s
```

[48]:
```
!head ./exp_2/dict_rules_bimmfallback_bimmboost150_ctest.seg
```

၁၉၆၂ ခုနှစ် ခန့်မှန်း သန်းခေါင်စာရင်း အရ လူဦး ရေ ၁၁၅၉၃၁ ယောက် ရှိ သည်
လူ တိုင်း တွင် သင့်မြတ် လျော်ကန် စွာ ကန့်သတ် ထား သည့် အလုပ် လုပ်ချိန် အပြင် လစာ
နှင့်တကွ အခါ ကာလ အားလျော်စွာ သတ်မှတ် ထား သည့် အလုပ် အားလပ်ရက် များ ပါဝင် သည့်
အနားယူခွင့် နှင့် အားလပ်ခွင့် ခံစားပိုင်ခွင့် ရှိ သည်
ဤ နည်း ကို စစ်ယူ သော နည်း ဟု ခေါ် သည်
စာပြန်ပွဲ ဆို တာ က အာဂုံဆောင် အလွတ်ကျက် ထား တဲ့ ပိဋကတ်သုံးပုံ စာပေ တွေ ကို စာစစ်
သံဃာတော်ကြီး တွေ ရဲ့ ရှေ့မှာ အလွတ် ပြန် ပြီး ရွတ်ပြ ရ တာ ပေါ့
ဒီ မှာ ကျွန်တော့် သက်သေခံကတ် ပါ
၂ ဝ ရာစု မြန်မာ့ သမိုင်း သန်း ဝင်း လှိုင် ၂ ဝ ဝ၉ ခု မေ လ ကံကော်ဝတ်ရည် စာပေ

ကျွန်တော် မျက်မှန် တစ် လက်လုပ် ချင် ပါ တယ်
ကျွန်တော် တို့ က ဒီ အမှု ရဲ့ ကြံရာပါ ကို ဖမ်းမိ ဖို့ ကြိုးစား ခဲ့ တယ်
ကလေး မီးဖွား ဖို့ ခန့်မှန်း ရက် က ဘယ်တော့ ပါ လဲ
အရိုးရှင်းဆုံး ကာဗိုဟိုက်ဒရိတ် မှာ ဂလူးကို့စ် ဂလက်တို့စ် ဖရုပ်တို့စ် စသည့်
မိုနိုဆက်ကရိုက် များ ဖြစ် သည်

[49]:
```
!python ./tools/eval_segmentation.py -r ./data/10k_test.txt -H ./exp_2/
↪dict_rules_bimmfallback_bimmboost150_ctest.seg
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.9034
Word Recall                    0.8807
Word F1-score                  0.8919
------------------------------------------------------------
Boundary Precision             0.6263
Boundary Recall                0.6106
Boundary F1-score              0.6184
------------------------------------------------------------
Vocab Precision                0.8245
Vocab Recall                   0.9208
Vocab F1-score                 0.8700
============================================================

Additional Statistics:
Reference words: 117857
Hypothesis words: 114896
Correct words: 103794
Reference vocabulary size: 10840
Hypothesis vocabulary size: 12107
Common vocabulary: 9982

Top Segmentation Errors Analysis
============================================================
Total errors: 41517

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):

Most Frequent Complex Boundary Errors:
  155 × REF: 'ပါ' → HYP: 'တယ်'
  100 × REF: 'ဖြစ်' → HYP: 'သည်'
   73 × REF: 'ခဲ့' → HYP: 'သည်'
```

```
59 × REF: 'ကြ' → HYP: 'သည်'
58 × REF: 'ခဲ့' → HYP: 'တယ်'
55 × REF: 'ပေး' → HYP: 'ပါ'
54 × REF: 'များ' → HYP: 'ကို'
52 × REF: 'ချင်' → HYP: 'ပါ'
52 × REF: 'သည်' → HYP: 'ခဲ့'
50 × REF: 'မ' → HYP: 'ရှိ'
```

0.9 with Bigger Dictionary, open-test

ဒီတခါတော့ myG2P+myPOS+Name Dictionary ကို သုံးမယ်။

[50]:
```
!time python oppa_word.py \
  --input "./data/otest.1k.word" \
  --output "exp_2/big_only_bimmfallback_bimmboost150_otest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos_name.dict" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.145s
user    0m0.128s
sys     0m0.016s
```

0.10 Evaluation

[51]:
```
!python ./tools/eval_segmentation.py -r ./data/otest.1k.word -H ./exp_2/
 ↪big_only_bimmfallback_bimmboost150_otest.seg --top-k 20
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.8813
Word Recall                    0.8388
Word F1-score                  0.8595
------------------------------------------------------------
Boundary Precision             0.5349
Boundary Recall                0.5091
Boundary F1-score              0.5216
------------------------------------------------------------
Vocab Precision                0.8036
Vocab Recall                   0.9151
Vocab F1-score                 0.8557
============================================================
```

```
Additional Statistics:
Reference words: 13468
Hypothesis words: 12818
Correct words: 11297
Reference vocabulary size: 2709
Hypothesis vocabulary size: 3085
Common vocabulary: 2479

Top Segmentation Errors Analysis
============================================================
Total errors: 5862

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):

Most Frequent Complex Boundary Errors:
   83 × REF: 'တယ်' → HYP: '။'
   74 × REF: 'ပါ' → HYP: '။'
   62 × REF: 'သည်' → HYP: '။'
   42 × REF: 'ပါ' → HYP: 'ပါတယ်'
   26 × REF: 'မယ်' → HYP: '။'
   19 × REF: 'ဘူး' → HYP: '။'
   18 × REF: 'လဲ' → HYP: '။'
   18 × REF: 'ဖြစ်' → HYP: 'သည်'
   17 × REF: 'လား' → HYP: '။'
   17 × REF: 'ခဲ့' → HYP: 'သည်'
   15 × REF: '။' → HYP: 'သည်'
   13 × REF: 'ဒီ' → HYP: 'ဒီနေ့'
   12 × REF: 'မ' → HYP: 'မရှိ'
   12 × REF: 'သလဲ' → HYP: '။'
   11 × REF: 'တယ်' → HYP: '၊'
   10 × REF: 'ခဲ့' → HYP: 'တယ်'
   10 × REF: 'ရှိ' → HYP: '။'
   10 × REF: 'ပါ' → HYP: 'ပါသလဲ'
    9 × REF: 'ချင်' → HYP: 'ချင်တယ်'
    9 × REF: 'ရ' → HYP: '။'
```

0.11 with Bigger Dictionary+Rules, open-test

[52]:
```
!time python oppa_word.py \
  --input "./data/otest.1k.word" \
  --output "exp_2/big_rules_bimmfallback_bimmboost150_otest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos_name.dict" \
```

```
  --postrule-file "./data/rules.txt" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.143s
user    0m0.129s
sys     0m0.014s
```

0.12 Evaluation

[53]:
```
!python ./tools/eval_segmentation.py -r ./data/otest.1k.word -H ./exp_2/
 ↪big_rules_bimmfallback_bimmboost150_otest.seg --top-k 20
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.9055
Word Recall                    0.8829
Word F1-score                  0.8941
------------------------------------------------------------
Boundary Precision             0.5991
Boundary Recall                0.5841
Boundary F1-score              0.5915
------------------------------------------------------------
Vocab Precision                0.8078
Vocab Recall                   0.9125
Vocab F1-score                 0.8570
============================================================

Additional Statistics:
Reference words: 13468
Hypothesis words: 13132
Correct words: 11891
Reference vocabulary size: 2709
Hypothesis vocabulary size: 3060
Common vocabulary: 2472

Top Segmentation Errors Analysis
============================================================
Total errors: 5126

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):
```

```
Most Frequent Complex Boundary Errors:
   53 × REF: 'တယ်' → HYP: '။'
   52 × REF: 'သည်' → HYP: '။'
   38 × REF: 'ပါ' → HYP: '။'
   20 × REF: '။' → HYP: 'သည်'
   20 × REF: 'ပါ' → HYP: 'တယ်'
   20 × REF: 'မယ်' → HYP: '။'
   18 × REF: 'ဖြစ်' → HYP: 'သည်'
   17 × REF: 'လား' → HYP: '။'
   15 × REF: 'ဘူး' → HYP: '။'
   12 × REF: 'လဲ' → HYP: '။'
   10 × REF: 'ကြ' → HYP: 'သည်'
    9 × REF: 'ခဲ့' → HYP: 'တယ်'
    9 × REF: '။' → HYP: 'တယ်'
    9 × REF: 'ခဲ့' → HYP: 'သည်'
    8 × REF: 'ရ' → HYP: '။'
    8 × REF: 'သည်' → HYP: 'ခဲ့'
    8 × REF: 'ခဲ့' → HYP: '။'
    8 × REF: 'ချင်' → HYP: 'တယ်'
    7 × REF: 'တစ်' → HYP: 'တစ်ခုခု'
    7 × REF: 'သည်' → HYP: 'ဖြစ်'
```

0.13 Big Dictionary (Closed-Test)

[54]:
```
!time python oppa_word.py \
  --input "./data/10k_test.txt" \
  --output "exp_2/big_only_bimmfallback_bimmboost150_ctest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos_name.dict" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.675s
user    0m0.659s
sys     0m0.016s
```

0.14 Evaluation

[55]:
```
!python ./tools/eval_segmentation.py -r ./data/10k_test.txt -H ./exp_2/
 ↪big_only_bimmfallback_bimmboost150_ctest.seg --top-k 20
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
```

```
------------------------------------------------------------
Word Precision                0.8753
Word Recall                   0.8311
Word F1-score                 0.8526
------------------------------------------------------------
Boundary Precision            0.5545
Boundary Recall               0.5265
Boundary F1-score             0.5401
------------------------------------------------------------
Vocab Precision               0.8097
Vocab Recall                  0.9221
Vocab F1-score                0.8623
============================================================

Additional Statistics:
Reference words: 117857
Hypothesis words: 111909
Correct words: 97949
Reference vocabulary size: 10840
Hypothesis vocabulary size: 12345
Common vocabulary: 9996

Top Segmentation Errors Analysis
============================================================
Total errors: 48805

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):

Most Frequent Complex Boundary Errors:
  421 × REF: 'ပါ' → HYP: 'ပါတယ်'
  117 × REF: 'ဒီ' → HYP: 'ဒီနေ့'
  114 × REF: 'ဖြစ်' → HYP: 'သည်'
  106 × REF: 'မ' → HYP: 'မရှိ'
   97 × REF: 'ခဲ့' → HYP: 'သည်'
   88 × REF: 'သူ' → HYP: 'သူက'
   79 × REF: 'မ' → HYP: 'မဟုတ်'
   71 × REF: 'က' → HYP: 'ကတော့'
   70 × REF: 'ကြ' → HYP: 'သည်'
   68 × REF: 'ခဲ့' → HYP: 'တယ်'
   67 × REF: 'ရှိ' → HYP: 'ရှိတယ်'
   65 × REF: 'များ' → HYP: 'ကို'
   61 × REF: 'ချင်' → HYP: 'ချင်တယ်'
   56 × REF: 'ပေး' → HYP: 'ပါ'
```

```
56 × REF: 'မြန်မာ' → HYP: 'မြန်မာနိုင်ငံ'
54 × REF: 'တို့' → HYP: 'တို့၏'
53 × REF: 'ချင်' → HYP: 'ချင်ပါတယ်'
52 × REF: 'ပါ' → HYP: 'ဘူး'
51 × REF: 'မ' → HYP: 'ဘူး'
50 × REF: 'မှာ' → HYP: 'မှာရှိ'
```

0.15 Big Dictionary+Rules (Closed-Test)

[57]:
```
!time python oppa_word.py \
  --input "./data/10k_test.txt" \
  --output "exp_2/big_rules_bimmfallback_bimmboost150_ctest.seg" \
  --space-remove-mode "my_not_num" \
  --dict "./data/myg2p_mypos_name.dict" \
  --postrule-file "./data/rules.txt" \
  --use-bimm-fallback \
  --bimm-boost 150
```

```
real    0m0.706s
user    0m0.685s
sys     0m0.014s
```

0.16 Evaluation

[58]:
```
!python ./tools/eval_segmentation.py -r ./data/10k_test.txt -H ./exp_2/
↪big_rules_bimmfallback_bimmboost150_ctest.seg --top-k 20
```

```
Word Segmentation Evaluation Results
============================================================
Metric                          Score
------------------------------------------------------------
Word Precision                 0.8993
Word Recall                    0.8739
Word F1-score                  0.8864
------------------------------------------------------------
Boundary Precision             0.6158
Boundary Recall                0.5984
Boundary F1-score              0.6070
------------------------------------------------------------
Vocab Precision                0.8100
Vocab Recall                   0.9202
Vocab F1-score                 0.8616
============================================================

Additional Statistics:
Reference words: 117857
```

```
Hypothesis words: 114527
Correct words: 102990
Reference vocabulary size: 10840
Hypothesis vocabulary size: 12315
Common vocabulary: 9975

Top Segmentation Errors Analysis
============================================================
Total errors: 42691

Most Frequent Over-Segmentation Errors (System split where it shouldn't):

Most Frequent Under-Segmentation Errors (System joined what should be separate):

Most Frequent Complex Boundary Errors:
  163 × REF: 'ပါ' → HYP: 'တယ်'
  100 × REF: 'ဖြစ်' → HYP: 'သည်'
   83 × REF: 'ခဲ့' → HYP: 'သည်'
   69 × REF: 'ကြ' → HYP: 'သည်'
   60 × REF: 'ခဲ့' → HYP: 'တယ်'
   58 × REF: 'ပေး' → HYP: 'ပါ'
   58 × REF: 'များ' → HYP: 'ကို'
   52 × REF: 'ချင်' → HYP: 'ပါ'
   52 × REF: 'မ' → HYP: 'ရှိ'
   52 × REF: 'သည်' → HYP: 'ခဲ့'
   49 × REF: 'ချင်' → HYP: 'တယ်'
   47 × REF: 'နေ' → HYP: 'တယ်'
   44 × REF: 'နိုင်' → HYP: 'မလား'
   43 × REF: 'ပေါ်' → HYP: 'ပေါ်မှာ'
   43 × REF: 'လို့' → HYP: 'ရ'
   42 × REF: 'ပါ' → HYP: 'ဘူး'
   41 × REF: 'တယ်' → HYP: 'ပါ'
   37 × REF: 'ရှိ' → HYP: 'သည်'
   36 × REF: 'လေ' → HYP: 'သည်'
   36 × REF: 'တစ်' → HYP: 'ခု'
```

0.17 Analysis

0.18 Word Metrics

Experiment	Precision	Recall	F1-score
Dictionary-1 (Open-test)	0.8857	0.8454	0.8651
Dictionary-1+Post Rules (Open-test)	0.9094	0.8893	0.8992
Dictionary-1 (Closed-test)	0.8795	0.8378	0.8582

Experiment	Precision	Recall	F1-score
Dictionary-1+Post Rules (Closed-test)	0.9034	0.8807	0.8919

0.19 Boundary Metrics

Experiment	Precision	Recall	F1-score
Dictionary-1 (Open-test)	0.5516	0.5265	0.5387
Dictionary-1+Post Rules (Open-test)	0.6116	0.5981	0.6048
Dictionary-1 (Closed-test)	0.5635	0.5368	0.5498
Dictionary-1+Post Rules (Closed-test)	0.6263	0.6106	0.6184

0.20 Vocabulary Metrics

Experiment	Precision	Recall	F1-score
Dictionary-1 (Open-test)	0.8139	0.9184	0.8630
Dictionary-1+Post Rules (Open-test)	0.8183	0.9158	0.8643
Dictionary-1 (Closed-test)	0.8242	0.9228	0.8707
Dictionary-1+Post Rules (Closed-test)	0.8245	0.9208	0.8700

0.21 for Bigger Dictionary

0.22 Word Metrics

Experiment	Precision	Recall	F1-score
Big Dictionary (Open-test)	0.8813	0.8388	0.8595
Big Dictionary+Post Rules (Open-test)	0.9055	0.8829	0.8941
Big Dictionary (Closed-test)	0.8753	0.8311	0.8526
Big Dictionary+Post Rules (Closed-test)	0.8993	0.8739	0.8864

0.23 Boundary Metrics

Experiment	Precision	Recall	F1-score
Big Dictionary (Open-test)	0.5349	0.5091	0.5216
Big Dictionary+Post Rules (Open-test)	0.5991	0.5841	0.5915
Big Dictionary (Closed-test)	0.5545	0.5265	0.5401
Big Dictionary+Post Rules (Closed-test)	0.6158	0.5984	0.6070

0.24 Vocabulary Metrics

Experiment	Precision	Recall	F1-score
Big Dictionary (Open-test)	0.8036	0.9151	0.8557
Big Dictionary+Post Rules (Open-test)	0.8078	0.9125	0.8570
Big Dictionary (Closed-test)	0.8097	0.9221	0.8623
Big Dictionary+Post Rules (Closed-test)	0.8100	0.9202	0.8616

0.25 To Do

- myNER ထဲက PER, LOC, ORG တွေကို ဆွဲထုတ်ပြီး dictionary မှာ ထပ်ဖြည့်ရန်
- လူနာမည်၊ မြို့ရွာနာမည်၊ အဖွဲ့အစည်းနာမည်တွေပါတဲ့ ဒေတာကို ပြင်ဆင်ပြီး oppaWord ကို evaluation လုပ်ရန်။ အဲဒါဆိုရင် ရလဒ်က ပိုပြတ်ပြတ်သားသား မြင်ရမယ်လို့ မျှော်လင့်တယ်။

[ ]: