

G2P with CRF (Tutorial)

Ye Kyaw Thu

Visiting Researcher, Waseda University, Japan

Visiting Professor, NECTEC, Thailand

TOC

- Introduction to G2P
- Preprocessing
- G2P Modeling with CRFSuite
- Testing
- Evaluation
- Analysis on Current Tagging Errors
- Practical Exercise with “_”
- Published works for Myanmar language G2P
- Assignment: Sentence Level G2P

Introduction to G2P

- G2P (Grapheme to Phoneme)
- Conversion of Myanmar text into correct pronunciation
- Word level conversion examples

| | | | | | |
|---------|---|-----|----------|------|-------------------|
| ကဗျာ | က | ဗျာ | ga - | bja | |
| ကဗျာစပ် | က | ဗျာ | စပ် | ga - | bja sa' |
| ကဗျာဆရာ | က | ဗျာ | <u>ဆ</u> | ရာ | ga - bja hsa - ja |
| ကဗျာဆန် | က | ဗျာ | ဆန် | | ga - bja hsan |

Introduction to G2P

- Word level G2P examples with some grammatical information:

N + N

ဆေး + ခန်း (အခန်း) ----> ဆေးဂန်း

ဈေး + တန်း (အတန်း) ----> ဈေးဒန်း

မီး + ခိုး (အခိုး) ----> မီးဂိုး

မဲဇလီ + ဖူး (အဖူး) ----> မဲဇလီဗူး

ယဉ် + ကြော (အကြော) ----> ယဉ်ဂျော

ရင် + ခေါင်း (အခေါင်း) ----> ရင်ဂေါင်း

အိမ်ကြို့ + အိမ်ကြား (အကြား) ----> အိမ်ဂျို့ အိမ်ဂျား

Introduction to G2P

- Word level G2P examples with some grammatical information:

N + V + N

ဖင် + ထိုင် + ခုံ ----> ဖင်ဒိုင်ခုံ

ရေ + ကူး + ကန် ----> ရေကူးကန်

ကုန် + တင် + ကား ----> ကုန်တင်ကား

V + V + N

နေ + ချင် + စိတ် ----> နေချင်စိတ်

မြင် + ခဲ့ + တာ ----> မြင်ခဲ့တာ

ပြော + ချင် + တာ ----> ပြောချင်တာ

Introduction to G2P

- Many patterns exist and some of them are:

$[N+V_f]^{cl\ N\ ph}$, $[N+V_s]^{cl\ N\ ph}$, $[for\ N+V_s]^{cl\ N\ ph}$,

$[N+V]^{cl\ N\ ph}$, $[V+N_{cls}/\ for\ N]^{cl\ N\ ph}$, $[V+N]^{cl\ N\ ph}$,

$[N+N]^{cl\ N\ ph}$, $[N+N_{cls}/\ for\ N]^{cl\ N\ ph}$, $[N+N_{cls}]^{cl\ N\ ph}$,

$[N+V+N]^{cl\ N\ ph}$, $[N+N_{par}+N]^{cl\ N\ ph}$, $[N+N_{cls}+N_{par}]^{cl\ N\ ph}$,

$[N+N_{par}+V]^{cl\ N\ ph}$, $[N+V+V]^{cl\ N\ ph}$, $[N+V+V]^{cl\ N\ ph}$

(Thein Tun, Acoustic Phonetics and The Phonology of the Myanmar Language, 2nd Edition, pp. 226-227)

Introduction to G2P

- If the vowel combination of 1st syllable is "င/in/aun" or "ဉ/in" or "န/an/un" or "မ/an/ein" or "ဝ/e" or "ဲ/e:" or "ံ/an", and the consonant of 2nd syllable is an unaspirated or an aspirated consonant that is unvoiced, then that 2nd syllable's pronunciation is voiced. Example pronunciations of some words are as follows:

တောင်ဗြဲ: (taun pjoun:) => တောင်ဗြဲ: (taun bjoun:)
ပိုင်းဇြဲ (pain: chei) => ပိုင်းဇြဲ (pain: gyei)

(Ye Kyaw Thu et. al., Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion", In Proceedings of the 13th International Conference on Computer Applications (ICCA 2015), February 5~6, 2015, Yangon, Myanmar, pp. 161-167)

Introduction to G2P

- There are many exceptions:

ရင်ခွန် (j in khoun) => ရင်ခွန် (j in khoun)
စည်းကမ်း (si: kan:) => စည်းကမ်း (si: kan:)

(Ye Kyaw Thu et. al., Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion", In Proceedings of the 13th International Conference on Computer Applications (ICCA 2015), February 5~6, 2015, Yangon, Myanmar, pp. 161-167)

Introduction to G2P

- If the vowel combination of 1st syllable is "င်/in/aun" or "၌/in" or "န်/an/un" or "မ်/an/ein" or "ဝ်/e" or "ဲ/e:" or "ံ/an", and the consonant of 2nd syllable is an unaspirated or an aspirated consonant that is unvoiced, then that 2nd syllable's pronunciation is voiced. Example pronunciations of some words are as follows:

တောင်ပုံ: (taun pjoun:) => တောင်ပုံ: (taun bjoun:)
ပိုင်ခွဲ (pain: chei) => ပိုင်ဂျဲ (pain: gyei)

(Ye Kyaw Thu et. al., Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion", In Proceedings of the 13th International Conference on Computer Applications (ICCA 2015), February 5~6, 2015, Yangon, Myanmar, pp. 161-167)

Introduction to G2P

- Word level + Sentence level conversion examples

ရေအေးတယ် ----> ရေ အေး ဒယ်
သူ့ကိုပြောပေးပါ ----> သူ့ ကို ပြော ပေး ဘာ
ကလေးတွေသောင်းကျန်းနေပြီ ----> က လေး တွေ သောင်း ကျန်း နေ ဘီ
နီးနီးကြားကြားရှိကြပါ ----> နီး နီး ကြား ရှား ရှိ ရ ဘာ
သိသိချင်းတားပါသေးတယ် ----> သိ သိ ချင်း တား ဘာ သေး ဒယ်

(Thein Tun, Acoustic Phonetics and The Phonology of the Myanmar Language, 2nd Edition, pp. 290, 294, 297)

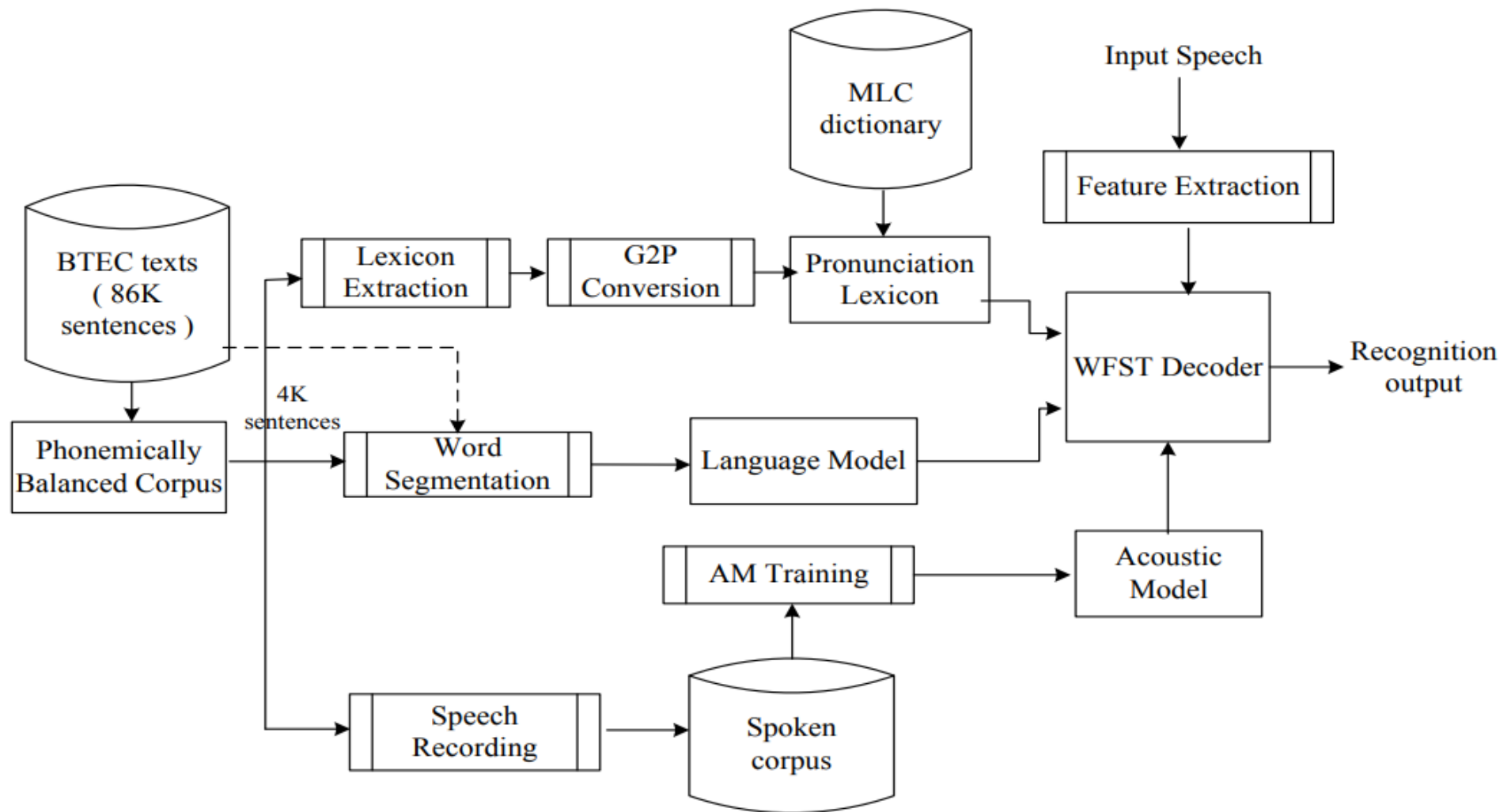
- Note: G2P is **not only for word level**, we have to prepare for **sentence level conversions**

G2P Dictionary

- Knowing how words are pronounced is an essential for building automatic speech recognition (ASR) and text-to-speech (TTS) systems
- Middle layer:
ASR Model -- → G2P Model -- → TTS Model
- Several approaches already proposed
- How about doing G2P conversion for Myanmar language?

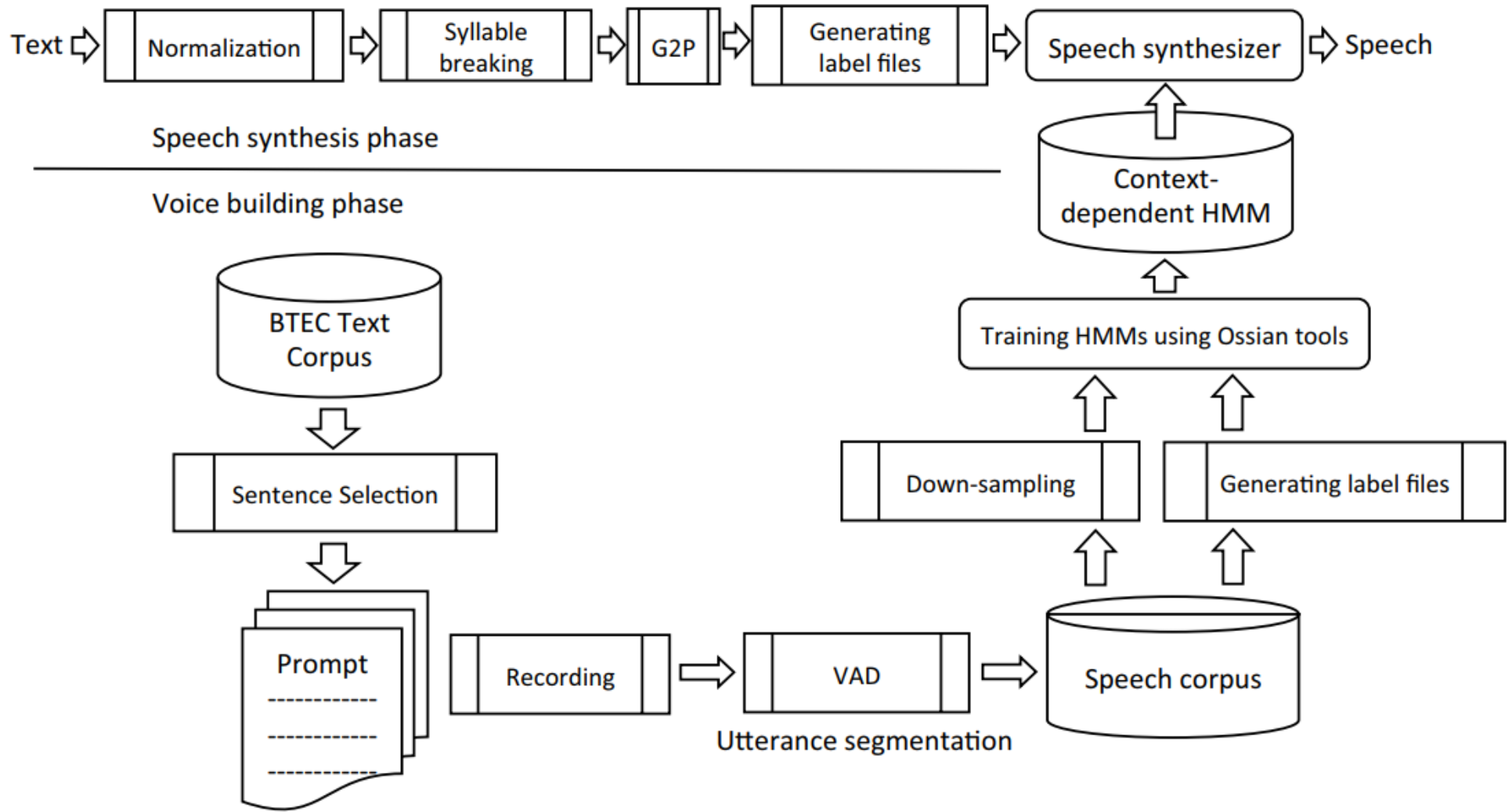
Introduction to G2P

- An Example of using G2P for Myanmar ASR



Introduction to G2P

- An example of using G2P for Myanmar TTS



Introduction to G2P

- Grapheme to phoneme mapping for consonants

ဇယားနံပါတ်(၁) - ဗျည်း အုပ်စုများ နှင့် သူတို့၏ အသံထွက်များ

| Grouped consonants | | | | |
|--------------------|--------------|--------|-------|----------|
| Unaspirated | Aspirated | Voiced | | Nasal |
| က /k/ | ခ /kh/ | ဂ /g/ | ဃ /g/ | င /ng/ |
| စ /s/ | ဆ /hs/ | ဇ /z/ | ဈ /z/ | ည/ည /nj/ |
| တူ /t/ | တူ /ht/ | သု /d/ | ဗ /d/ | ဏ /n/ |
| တ /t/ | တ /ht/ | ဒ /d/ | ဓ /d/ | န /n/ |
| ပ /p/ | ဖ /hp/ | ဗ /b/ | ဘ /b/ | မ /m/ |
| ယ /j/ | ရ /j/ or /r/ | လ /l/ | ဝ /w/ | သ /th/ |
| | ဟ /h/ | ဌ /l/ | အ /a/ | |

- Ref: <https://github.com/ye-kyaw-thu/myG2P>

Introduction to G2P

- Some examples of vowel combinations and their pronunciations

ဇယားနံပါတ်(၂) - "အ" ဗျည်း နှင့် သရ ပေါင်းစပ်မှု ပုံစံများ နှင့် သူတို့၏ အသံထွက်များ

| | | | | | | |
|--------|---------|---------|-----------|-----------|------------|-------------|
| အိ i | အိ i. | အိး i: | အိ' i' | အင် in | အင် in. | အင်း in: |
| အေ ei | အေ့ ei. | အေး ei: | အိတ် ei' | အိန် ein | အိန့် ein. | အိန်း ein: |
| အယ် e | အယ့် e. | အဲ e: | အိုက် ai' | အိုင် ain | အိုင် ain. | အိုင်း ain: |
| အာ a | အာ့ a. | အား a: | အတ် a' | အန် an | အန့် an. | အန်း an: |
| အော် o | အော့ o. | အော o: | အောက် au' | အောင် aun | အောင် aun. | အောင်း aun: |
| အု u | အု u. | အူး u: | အုတ် u' | အုန် un | အုန့် un. | အုန်း un: |
| အံ ou | အံ့ ou. | အံး ou: | အုပ် ou' | အုန် oun | အုန့် oun. | အုန်း oun: |

Introduction to G2P

- G2P is a very good Lab. exercise for Myanmar language NLP
- We will use one of the sequence modelings, CRF (Conditional Random Fields)
- Toolkits: I recommend to use CRF++ or CRFSuite
- For this exercise we will use CRFSuite
- Refer following CRFSuite Tutorial:
<http://www.chokkan.org/software/crfsuite/tutorial.html>

G2P Dictionary

- We will use myG2P Dictionary (Version 1) for this Lab exercise

<https://github.com/ye-kyaw-thu/myG2P/tree/master/ver1>



Ye Kyaw Thu

ye-kyaw-thu

myG2P

Myanmar (Burmese) Language Grapheme to Phoneme (myG2P) Conversion Dictionary for speech recognition (ASR) and speech synthesis (TTS).

grapheme-to-phone

g2p

myanmar

dictionary

myanmar-grapheme

★ 15

🔗 2

Updated on Jul 29, 2017

Preprocessing (Installation of CRFSuite)

- Link of CRFSuite:

<http://www.chokkan.org/software/crfsuite/>

- After you download CRFSuite, untar:

```
$ tar -xzvf ./crfsuite-0.12.tar.gz
```

Preprocessing (Installation of CRFSuite)

- `$./autogen.sh`

libtoolize: error: One of these is required:

libtoolize: gm4 gnum4 m4

libtoolize: error: Please install GNU M4, or 'export M4=/path/to/gnu/m4'.

`./autogen.sh: 20: ./autogen.sh: aclocal: not found`

aclocal failed!

- If you got above error, you need to install automake:

`$ sudo apt-get install automake`

Preprocessing (Installation of CRFSuite)

- After running “autogen.sh” successfully
- Run as usual:
 - ./configure
 - make
 - sudo make install
- You might need to run ./configure with some options
 - \$./configure --with-liblbfgs=/home/yekyawthu/tool/liblbfgs-1.10

Preprocessing (Installation of CRFSuite)

- After installation, when you run “crfsuite –help”, if you got following error:
 - crfsuite: error while loading shared libraries:
libcrfsuite-0.12.so: cannot open shared object file: No such file or directory
- You might need to update “ld.so.conf” file:
 - \$ sudo vi /etc/ld.so.conf
- I added following line:
 - LIBDIR=/home/yekyawthu/local/lib/

Preprocessing (Installation of CRFSuite)

- If you installed successfully

```
lar@lar-air:~$ crfsuite -h
CRFSuite 0.12 Copyright (c) 2007-2011 Naoaki Okazaki

USAGE: crfsuite <COMMAND> [OPTIONS]
      COMMAND      Command name to specify the processing
      OPTIONS      Arguments for the command (optional; command-specific)

COMMAND:
  learn           Obtain a model from a training set of instances
  tag             Assign suitable labels to given instances by using a model
  dump           Output a model in a plain-text format

For the usage of each command, specify -h option in the command argument.
lar@lar-air:~$
```

Preprocessing (Cloning myG2P dictionary)

- You can get the myG2P data with following command:

```
git clone https://github.com/ye-kyaw-thu/myG2P
```

- Dictionary is under following path:

```
./myG2P/ver1/
```

- Dictionary filename: myg2p.ver1.txt

Preprocessing (Check myG2p dictionary)

```
$ head ./myg2p.ver1.txt
```

```
1    ...ဖြစ်စေ...ဖြစ်စေ    ... ဖြစ် စေ... ဖြစ် စေ    ... hpji' sei ... hpji' sei
2    ...ရို...စဉ်    ... ရို.. စဉ်    jou: ... sin
3    ...ရို...စဉ်    ... ရို.. စဉ်    jou: ... zin
4    ...လို...ငြ    ... လို... ငြ    ... lou ... nja:
5    ကကတစ်    က က တစ် ka. ga- di'
6    ကကတို    က က တို ka. ga- dou:
7    ကကုသန်    က ကု သန် ka. ku. than
8    ကကုသန်    က ကု သန် kau' ka- than
9    ကကူရုံ    က ကူ ရုံ ka. ku jan
10   ကကြို    က ကြို ka. gyou:
```

- We should remove top 4 lines from the original myG2p dictionary

Preprocessing (Cutting only column-3 and 4)

- `$cut -f2,3 ./myg2p.ver1.txt > f23`

- `$ head f34`

က က တစ် ka. ga- di'

က က တို့ ka. ga- dou:

က ကု သန် ka. ku. than

က ကု သန် kau' ka- than

က ကူ ရံ ka. ku jan

က ကြံ ka. gyau:

က ကြံ တန် ဆ ka. gyau: da- za

က ကြံ က ကြောင် လုပ် ga- gyi ga- gyaun lou'

က ကြံ ka. gyi:

က ကြံ ထွန် ka. gyi: htun

Preprocessing

(Dividing training and test data)

- Use “shuf” command for shuffling
`$shuf f34 > f34.shuf`
- Use “head” and “tail” commands for splitting training and test data (80% for training and 20% for testing)

```
$head -n 19830 ./f34.shuf > ./myg2p.dictver1.train
```

```
$tail -n 4957 ./f34.shuf > ./myg2p.dictver1.test
```

- Check no. of lines with “wc” command:

```
$wc myg2p.dictver1.train
```

```
19830 56820 867421 myg2p.dictver1.train
```

```
$wc myg2p.dictver1.test
```

```
4957 14199 216868 myg2p.dictver1.test
```

Preprocessing (grapheme phoneme y format)

- For training data

```
$perl ch2col2.pl ./myg2p.dictver1.train > ./myg2p.dictver1.train.col
```

- For testing data

```
$perl ch2col2.pl ./myg2p.dictver1.test > ./myg2p.dictver1.test.col
```

- Check the format of training:

```
$head myg2p.dictver1.train.col
```

၀ sa- sa-

ရွှေ jwei_ jwei_

တံ dan dan

ကြွေ kywe_ kywe_

မြီ mi_ mi_

တနံ tan_ tan_

သ tha- tha-

မလံ me' me'

Preprocessing (Change to CRFSuite format)

- For Training

```
$cat myg2p.dictver1.train.col | perl ./chunking.py > train
```

- For Testing

```
$cat myg2p.dictver1.test.col | perl ./chunking.py > test
```

Preprocessing (Change to CRFSuite format)

- When you check the format with “head” command:

```
tha-   w[0]=သု w[1]=ငယ်   w[2]=မ   w[0]|w[1]=သုငယ် pos[0]=tha-   pos[1]=nge   pos[2]=ma.   pos[0]|pos[1]=tha-|ng
e      pos[1]|pos[2]=nge|ma.   pos[0]|pos[1]|pos[2]=tha-|nge|ma.   __BOS__
nge    w[-1]=သု w[0]=ငယ်   w[1]=မ   w[-1]|w[0]=သုငယ်   w[0]|w[1]=ငယ်မ pos[-1]=tha-   pos[0]=nge   pos[1]=ma.pos[
-1]|pos[0]=tha-|nge pos[0]|pos[1]=nge|ma.   pos[-1]|pos[0]|pos[1]=tha-|nge|ma.
ma.    w[-2]=သု w[-1]=ငယ်   w[0]=မ   w[-1]|w[0]=ငယ်မ   pos[-2]=tha-   pos[-1]=nge   pos[0]=ma.   pos[-2]|pos[-
1]=tha-|nge pos[-1]|pos[0]=nge|ma.   pos[-2]|pos[-1]|pos[0]=tha-|nge|ma.   __EOS__

lou'   w[0]=လူ   w[1]=ပိုင်   w[2]=ခင်   w[0]|w[1]=လူပိုင် pos[0]=lou'   pos[1]=pain   pos[2]=gwin.   p
os[0]|pos[1]=lou'|pain pos[1]|pos[2]=pain|gwin.   pos[0]|pos[1]|pos[2]=lou'|pain|gwin.   __BOS__
pain   w[-1]=လူ   w[0]=ပိုင်   w[1]=ခင်   w[-1]|w[0]=လူပိုင်   w[0]|w[1]=ပိုင်ခင် pos[-1]=lou'   pos[0]=painp
os[1]=gwin.   pos[-1]|pos[0]=lou'|pain   pos[0]|pos[1]=pain|gwin.   pos[-1]|pos[0]|pos[1]=lou'|pain|gwin.
gwin.   w[-2]=လူ   w[-1]=ပိုင်   w[0]=ခင်   w[-1]|w[0]=ပိုင်ခင်   pos[-2]=lou'   pos[-1]=pain   pos[0]=gwin.   p
os[-2]|pos[-1]=lou'|pain   pos[-1]|pos[0]=pain|gwin.   pos[-2]|pos[-1]|pos[0]=lou'|pain|gwin.   __EOS__

kyain: w[0]=ကျိန်   w[1]=ခေတ်   w[2]=မူ   w[0]|w[1]=ကျိန်ခေတ် pos[0]=kyain__COLON__   pos[1]=gaun__COLON__   p
os[2]=mwei   pos[0]|pos[1]=kyain__COLON__|gaun__COLON__   pos[1]|pos[2]=gaun__COLON__|mwei   pos[0]|pos[1]|pos[2]
=kyain__COLON__|gaun__COLON__|mwei   __BOS__
gaun:   w[-1]=ကျိန်   w[0]=ခေတ်   w[1]=မူ   w[-1]|w[0]=ကျိန်ခေတ်   w[0]|w[1]=ခေတ်မူ pos[-1]=kyain__COLON__   p
os[0]=gaun__COLON__   pos[1]=mwei   pos[-1]|pos[0]=kyain__COLON__|gaun__COLON__   pos[0]|pos[1]=gaun__COLON__|mwei   p
os[-1]|pos[0]|pos[1]=kyain__COLON__|gaun__COLON__|mwei
```

G2P Modeling with CRFSuite

- “crfsuite learn” command is for modeling or training
- “crfsuite tag” command is for tagging or testing
- Check the option with following commands:
 \$crfsuite learn --help
 \$crfsuite tag --help

G2P Modeling with CRFSuite

- Building g2p model with following command:

```
$crfsuite learn -m ./g2p.model -e2 ./train ./test
```

- Here,

- m, --model=FILE for output model name

- Note: By default, this utility does not store the model and thus you should use “-m or –model” option

- e2 option performs a holdout evaluation on the data set #2 (i.e. ./test)

G2P Modeling with CRFSuite

- When you run:
\$crfsuite learn
-m ./g2p.model
-e2 ./train ./test
- Training time will be depends on
your data size,
no. of features
and learning
options

```
CRFSuite 0.12 Copyright (c) 2007-2011 Naoaki Okazaki
Start time of the training: 2019-01-07T11:51:16Z

Reading the data set(s)
[1] ./train
0....1....2....3....4....5....6....7....8....9....10
Number of instances: 19831
Seconds required: 0.558
[2] ./test
0....1....2....3....4....5....6....7....8....9....10
Number of instances: 4958
Seconds required: 0.155

Statistics the data set(s)
Number of data sets (groups): 2
Number of instances: 24787
Number of items: 71019
Number of attributes: 221448
Number of labels: 1868

Holdout group: 2

Feature generation
type: CRF1d
feature.minfreq: 0.000000
feature.possible_states: 0
feature.possible_transitions: 0
0....1....2....3....4....5....6....7....8....9....10
Number of features: 362629
Seconds required: 0.419

L-BFGS optimization
c1: 0.000000
c2: 1.000000
num_memories: 6
max_iterations: 2147483647
epsilon: 0.000010
stop: 10
delta: 0.000010
linesearch: MoreThuente
linesearch.max_iterations: 20
```


Testing

- Run following command for testing:
`$crfsuite tag -m ./g2p.model ./test | tee test1.out`
- Here,
-m or --model option is for the model name that you built
“tee” command is for reading standard input and write to standard output and files

Testing

- You will get predicted “phonemes” or “y” as one column:

```
pjei.  
mjau'  
  
taun  
jou  
  
ta-  
ja  
  
bwa  
ga-  
ne  
  
mje'  
nha-  
nge
```

- And thus, you have to make parallel data with grapheme parts for evaluation process

Evaluation

(preprocessing for evaluation)

- 1st Change column to line with “ch2line.pl”
`$perl ./ch2line.pl ./test1.out > ./test1.out.line`
- Check the output with `$head ./test1.out.line`:

```
yekyawthu@bit-MS-7B09:~/exp/g2p/word$ head ./test1.out.line
tha- nge ma.
lou' pain gwin.
lhain gaun mwei
na- mji gwin
na' tha'
lun bjan
mwei mju jei
mje' nha- bei
khaun laun ti
kywe' kya.
```

Evaluation

(Preprocessing for evaluation)

```
$perl ./mk-wordtag.pl ./myg2p.dictver1.test "\V" w  
> ./myg2p.dictver1.test.word
```

```
$head ./myg2p.dictver1.test.word
```

သူ ငယ် မ
လုပ် ပိုင် ခွင့်
ကျိုင်း ခေါင်း မြေ
နွား မြီး ကွင်း
နတ် သတ်
လွန် ပြန်
မွေး မြူ ရေး
မျက် နှာ ပေး
ခေါင်း လောင်း တီး
ကြွက် ကျ

Evaluation

(Preprocessing for evaluation)

```
$perl ./mk-pair.pl ./myg2p.dictver1.test.word  
./test1.out.line > test1.out.line.wordtag
```

```
$head ./test1.out.line.wordtag
```

```
သူ/tha- cဝ်/nge မ/ma.  
လုပ်/lou' ပိုင်/pain ခွင့်/gwin.  
ကျိုင်း/lhain ခေါင်း/gaun မြေ/mwei  
နား/na- မြီး/mji ကွင်း/gwin  
နတ်/na' သတ်/tha'  
လွန်း/lun ပြန်/bjan  
မွေး/mwei မြူ/mju ရေး/jei  
မျက်/mje' နှာ/nha- ဝေး/bei  
ခေါင်း/khaun လောင်း/laun တီး/ti  
ကြွက်/kywe' ကျ/kya.
```

Evaluation

(Measuring tagging accuracy%)

```
$perl ./gradeupos.pl ./myg2p.dictver1.test  
./test1.out.wordtag
```

Accuracy: 72.67% (10318/14199)

Most common mistakes:

| | |
|----------------------------|----|
| မီး:/mi: --> မီး:/mi | 66 |
| လုံး:/loun: --> လုံး:/loun | 50 |
| ကြီး:/gyi: --> ကြီး:/gyi | 47 |
| ရေး:/jei: --> ရေး:/jei | 45 |
| လေး:/lei: --> လေး:/lei | 38 |
| ရိုး:/jou: --> ရိုး:/jou | 36 |
| ဦး:/u: --> ဦး:/u | 35 |
| စား:/za: --> စား:/za | 33 |
| သား:/dha: --> သား:/dha | 33 |
| စည်း:/si: --> စည်း:/si | 33 |

Analysis on Current Tagging Errors

- Currently we only got 72.67%
- Most of the errors are relating to Myanmar tone character Visarga or WitSaNhaLone Pauk

Accuracy: 72.67% (10318/14199)

Most common mistakes:

မီး:/mi: --> မီး:/mi 66
လုံ:/loun: --> လုံ:/loun 50
ကြီး:/gyi: --> ကြီး:/gyi 47
ရေး:/jei: --> ရေး:/jei 45
လေး:/lei: --> လေး:/lei 38
ရိုး:/jou: --> ရိုး:/jou 36
ဦး:/u: --> ဦး:/u 35
စား:/za: --> စား:/za 33
သား:/dha: --> သား:/dha 33
စည်း:/si: --> စည်း:/si 33

Practical Exercise with “_”

- It is strange and I think error occurred because of using colon character “:”
- When I run “head command on training data”:

```
$ head -5 train
```

```
sa- w[0]=၈ w[1]=၈: w[2]=၈ w[0]|w[1]=၈|၈: pos[0]=sa- pos  
[1]=jwei__COLON__ pos[2]=dan pos[0]|pos[1]=sa-|jwei__COLON__ pos[1]|pos  
[2]=jwei__COLON__|dan pos[0]|pos[1]|pos[2]=sa-|jwei__COLON__|dan __BOS__  
jwei: w[-1]=၈ w[0]=၈: w[1]=၈ w[-1]|w[0]=၈|၈: w[0]|w[1]=၈:|၈ pos[-1]=sa-  
pos[0]=jwei__COLON__ pos[1]=dan pos[-1]|pos[0]=sa-|jwei__COLON__ pos[0]|pos  
[1]=jwei__COLON__|dan pos[-1]|pos[0]|pos[1]=sa-|jwei__COLON__|dan  
dan w[-2]=၈ w[-1]=၈: w[0]=၈ w[-1]|w[0]=၈:|၈ pos[-2]=sa- pos  
[-1]=jwei__COLON__ pos[0]=dan pos[-2]|pos[-1]=sa-|jwei__COLON__ pos[-1]|pos  
[0]=jwei__COLON__|dan pos[-2]|pos[-1]|pos[0]=sa-|jwei__COLON__|dan __EOS__
```

```
kywe: w[0]=၈၈ w[1]=၈: w[2]=၈: w[0]|w[1]=၈၈|၈: pos[0]=kywe__COLON__  
pos[1]=mi__COLON__ pos[2]=tan__COLON__ pos[0]|pos[1]=kywe__COLON__|mi__COLON__  
pos[1]|pos[2]=mi__COLON__|tan__COLON__ pos[0]|pos[1]|pos[2]=kywe__COLON__|  
mi__COLON__|tan__COLON__ __BOS__
```


Practical Exercise with “_”

- How about replacing “:” symbol with “_”
- You can use sed or tr command

```
$cat ./myg2p.dictver1.train | tr ':' '_'
```

```
စ/sa-ရွေး/jwei_တံ/dan  
ကြွက်/kywe_မြီး/mi_တနီး/tan_  
သ/tha-မက်/me' ဖမ်း/hpan_  
ဖော့/hpo. ဆို့/zou.  
ခဲ/ge_ဖိုး/bou_  
လက်/le' စွပ်/su'  
စ/sa-မြင်/mj in  
ရာ/jaသိ/dhi ကုန်/goun  
လမ်း/lan_လျှောက်/shau' ဝင်/pin  
မာ/maသ/ga.
```

- Don't forget for test data also!

Practical Exercise with “_”

- Evaluation result with underscore model is 98.45%:

```
$ perl ./gradeupos.pl ./myg2p.dictver1.test ./test1.out.line.wordtag  
Accuracy: 98.45% (13979/14199)
```

Most common mistakes:

```
စာ/swa --> စာ/zwa    5  
ဇု/bju. --> ဇု/pju.   3  
ဟ/ha- --> ဟ/ha.    3  
ကွာ/gwa --> ကွာ/kwa   3  
ဆငံ/zin. --> ဆငံ/hsin. 2  
/shi' --> ဝံ/shi.    2  
/jwe. -->           2  
ရိတ်/rai' --> ရိတ်/jai' 2  
ဇု/pju --> ဇု/pju.    2  
ငွေ/gwei. --> ငွေ/ngwei. 2
```

Published Works for Myanmar Language G2P

- Word level G2P papers:

Ye Kyaw Thu, Win Pa Pa, Andrew Finch, Aye Mya Hlaing, Hay Mar Soe Naing, Eiichiro Sumita and Chiori Hori, "Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion", In Proceedings of the 13th International Conference on Computer Applications (ICCA 2015), February 5~6, 2015, Yangon, Myanmar, pp. 161-167. [Best Paper Award]

Ye Kyaw Thu, Win Pa Pa, Yoshinori Sagisaka, Naoto Iwahashi, "Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary", In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), COLING 2016, December 11-17, 2016, Osaka, Japan, pp. 11–22.

Published Works for Myanmar Language G2P

- Sentence level G2P papers:

Ye Kyaw Thu, Win Pa Pa, Andrew Finch, Jinfu Ni, Eiichiro Sumita and Chiori Hori, 2015, "The Application of Phrase Based Statistical Machine Translation Techniques to Myanmar Grapheme to Phoneme Conversion", In Proceedings of the Pacific Association for Computational Linguistics Conference (PACLING 2015), May 19~21, 2015, Legian, Bali, Indonesia, pp. 170-176. (revised paper has been published in Springer Communication in Computer and Information Science (CCIS), ISSN:1865-0929, pp. 238-250)

- Note: We used myG2P dictionary + extracted 5,276 sentences of BTEC corpus for this PACLING 2015 conference paper

Assignment: Sentence Level G2P

- 1st build a word-level G2P model
- Collect 1,000 Myanmar sentences from social media or web news domain
- Parse collected data with word level G2P model
- Manually fixed phoneme conversion errors of parsed sentences
- Shuffle your data and split (800 sentences for training and 200 sentences for testing)
- Build sentence-level G2P model with CRFSuite and make evaluation with 200 sentence test data
- Report the accuracy% of your sentence-level G2P in details

Myanmar Language G2P Research?!

- We need sentence level tagged corpus for several domains
- You can try with different modeling approaches to increase tagging accuracy %
- At first, try to finish my assignment
- You can also consider to add some more features (refer Ye Kyaw Thu et. al., ICCA2015)

| | | | | | |
|--------|---|-------|--------|---------|-------|
| ကောင်း | 0 | kaun: | ခံ | 0 0 0 c | ga- |
| ကောင်း | c | gaun: | တွင်း | c 0 0 c | dwin: |
| ကန်း | w | kan: | ကောင်း | w 0 0 0 | kaun: |
| ကန်း | c | gan: | | | |

Perl Scripts

- Following perl scripts uploaded to <https://github.com/ye-kyaw-thu/myG2P/tree/master/tutorial>
 1. ch2col2.pl
 2. mk-wordtag.pl
 3. mk-pair.pl
 4. ch2line.pl
 5. gradepos.pl