Exercise 13 G2P Tagging with rdr(Ripple Down Rule)

Phyo Thu Htet

4 Dec 2019 (Wed)

@Software Lab

#An exercise : a part of ASR

Burmese (Syllbreak Level)

#Grapping phoneme and grapheme

```
cut -f 3 myg2p.ver1.1.txt > BurmeseDictGrapheme
```

```
cut -f 4 myg2p.ver1.1.txt > Phoneme
```

#Word Count to check the relevancy

wc -l BurmeseDictGrapheme

wc -l BurmesePhoneme

#24798

Write a perl program for g2pmapping.(#to check which has no same length in syllable level)

Count:3841

G:စ ကားဖောင်

P:za- ga: hpaun

Count:24517G:

ဣန် ဒြေP:ein da- rei

Count:24518G:ဣန် ဒြေ ကြီးP:ein da- rei kyi:Count:24519G:ဣန် ဒြေ ဆည်P:ein da- rei hseCount:24520G:ဣန် ဒြေ ဆောင်P:ein da- rei hsaunCount:24521G:ဣန် ဒြေ ပျက်P:ein da- rei pje'Count:24522G:ဣန် ဒြေ မဲ့P:ein da- rei me.Count:24523G:ဣန် ဒြေ ရP:ein da- rei ja.Count:24524G:ဣန် ဒြေ ရှင်P:ein da- rei shinCount:24525G:ဣန် ဒြေ ရှိP:ein da- rei shi.Count:24526G:ဣန် ဒြေ လုပ်P:ein da- rei lou'Count:24527G:ဣန် ဒြေ သမ် ပတ် တိP:ein da- rei than pa' ti.

Manual Correction and Recheck

wc -l BurmeseDictGrapheme

wc -l BurmesePhoneme

#24798

Format

က/ka. က/ga- တစ်/di'

က/ka. က/ga- တိုး/dou:

က/ka. ကု/ku. သန်/than

က/kau' ကု/ka- သန်/than

က/ka. ကူ/ku ရုံ/jan

က/ka. ကြိုး/gyou:

English (Word Level)

The g2p dict file for English is used from here http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/cmudict-0.7b.

The format of the dict file (Sample)

E.g.

ACCEPTED **AE0 K S EH1 P T IH0 D**

```
# sed 's/\t/\//g' filename
```

```
# cut -f 2 filename
```

```
# sed 's/ /-/g' filename
```

```
# paste file1 file2
```

After that the format of the file would be like the following,

ACCEPTED **AE0-K-S-EH1-P-T-IH0-D**

Format

Model usage (rdr)

Why rdr?

It is fast.

How to proceed:

E.g.

```
pSCRDRtagger$ python RDRPOSTagger.py train ../data/goldTrain
```

```
pSCRDRtagger$ python RDRPOSTagger.py tag ../data/goldTrain.RDR ../data/goldTrain.DICT ../data/rawTest
```

Model Description

1. Only Burmese Data is used
2. Only English Data is used
3. **Both English and Burmese data are utilized**

The third one is considered cause the data that we will use for ASR model (Medical Domain) includes both language.

And for the sake of **Simplicity : The rdr provide one file for one model test.**

Data Testing

SyllBreak For Myanmar

Capitalize English Words

Result

E.g.

မ/ma- ကူး/ku: စက်/se' နိုင်/nain သော/tho: ရော/jo: ဂါ/ga များ/mja:

Future Work

Evaluation

Data Testing with word-syllable level for Myanamr

Testing with other approaches

Have a very nice day.

Tank you.