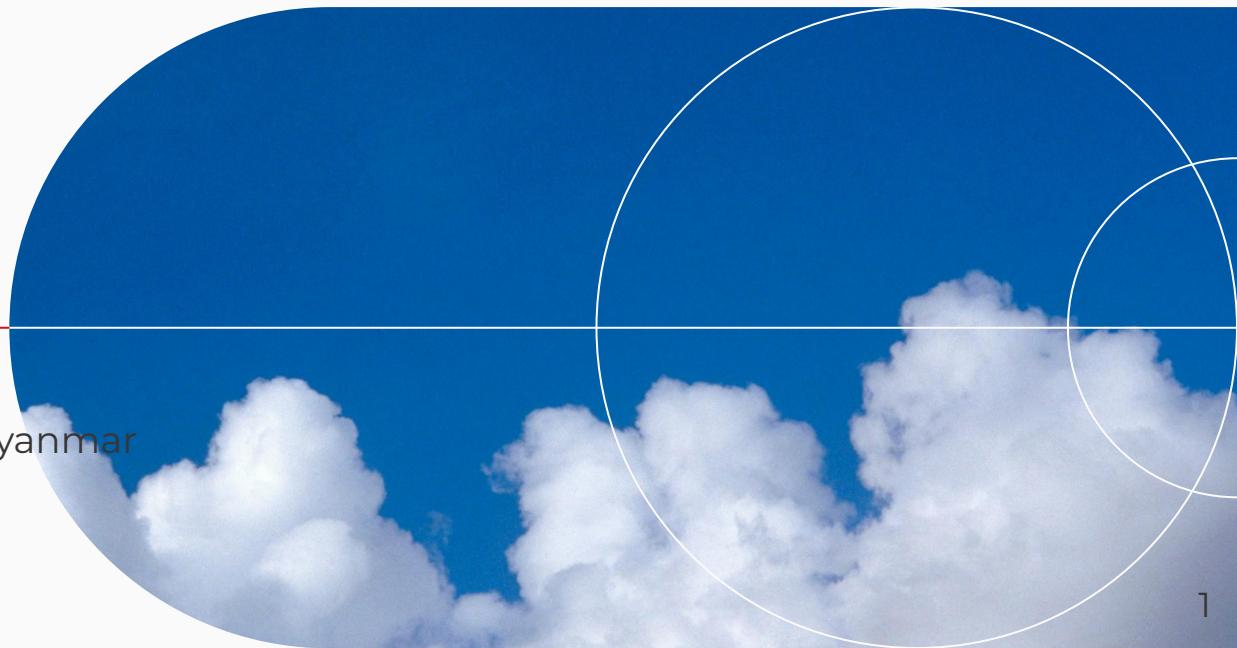


မြန်မာနှင့်ငံဘာသာစက္းနား အထွက် NLP လုပေသန

Ye Kyaw Thu

Lab. Leader,
Language Understanding Lab., Myanmar
Visiting Professor,
NECTEC, Thailand



Agenda

- 01** About Me
- 02** Corpora for NLP R&D
- 03** Embeddings
- 04** Language Model
- 05** Practical NLP Applications
- 06** Conclusion & Open Discussion

About Me



- Doctor of Science, GITS, Waseda University, Tokyo, Japan (26th October 2011)
- Master of Science, GITS, Waseda University, Tokyo, Japan (15th March 2006)
- Graduation of One Year Special Japanese Language Course, The Japanese Language School of the International Student Institute (国際学友会日本語学校), Tokyo, Japan (12th March 2004)
- Bachelor of Science (Physics), Dagon University, North Dagon, Myanmar (15th December 2000)
- International Advanced Diploma in Computer Studies, NCC Education, UK (13th December 1999)
- International Diploma in Computer Studies, NCC Education, UK (13th June 1998)

About Me



- Research Associate (Apr 2009 - Mar 2012) @Waseda University, Tokyo
- Researcher (Apr 2012 to Mar 2016)@NICT, Kyoto
- Researcher (Sept 2016 to Mar 2018)@Okayama Prefectural University

About Me



- Visiting Professor (from Jan 2019 to Present)
- Supervising students at various institutions, including Assumption University (AU), Kasetsart University (KU), King Mongkut's Institute of Technology Ladkrabang (KMITL), Sirindhorn International Institute of Technology (SIIT), Japan Advanced Institute of Science and Technology (JAIST)

About Me



- Lab Leader at LU Lab., Myanmar
- Established in 2019
- Internship Camp
- R&Ds with undergrad, master's and doctoral students

About Me

The screenshot shows the official website for the Multilingual Speech Translation Application "VoiceTra". The top navigation bar includes links for Home, Features, Manuals, FAQs, Contact, Videos and Screenshots, Terms of Use, and Report Errors. A purple banner at the top features a woman using a smartphone and the text "Communicate with the world using \"VoiceTra\"!". Below this is a large image of a smartphone displaying the app's user interface, which shows speech-to-speech translation between English and Japanese. A red button encourages users to "Download and Use for FREE!". At the bottom, there are download links for the App Store and Google Play, along with a "FLYER Download here" button and a video thumbnail titled "Play to learn about \"VoiceTra\"".

- VoiceTra Project
- <https://voicetra.nict.go.jp/en/>
- ဒီပရောဂုံးမှာ
မြန်မာစာကို
ထည့်နေဖော်အတွက်
ငြနစ်တတ် NICT
လုပေါ်တဲ့ ဂျပန်အစိုးရ^{၁၂}
သတေသနဌာန၊
ကျိုတိမှာ
အလုပလုပ်ခဲ့တယ်

Corpora for NLP R&D

- ခုက စပိုးဆောက်ခဲ့တဲ့ ကောပတ်စွဲတွေ အကြောင်းကို တစ်ခုပြီး တစ်ခု အကျဉ်းမံတ်ဆက် ပေးသွားပါမယ်
- **xCorpus**, **yCorpus**, **zCorpus** ... ဆိုပြီး
- မဟုတ်ဘူးယူ၊ တကယ်ကတော့ **myX**, **myY**, **myZ** ဆိုတဲ့ ပုံစံ ပိုများလိမ့်မယ်
- တကယ်ကို အမှန်းကြားရပါလိမ့်မယ် :)

Corpora (myG2P)

- We developed this **myG2P** (Myanmar Grapheme-to-Phoneme) dictionary for VoiceTra (Multilingual Speech Translation Application) Myanmar language project of **NICT, Japan (during 2014-2015)**.
- We mainly used **MLC (Myanmar Language Commission)** dictionary words.
- Please cite the **ICCA 2015** paper and/or **COLING 2016** paper, if you use myG2P dictionary.
- Please cite **PACLING 2015** paper, if you are talking about sentence level grapheme-to-phoneme conversion of Myanmar language.

Corpora (myG2P)

- တံခါး (tan kha:) ==> da- ga:
- နာမည်ကောင်း (na mji kaun:) ==> na mji gaun:, nan me gaun:, nan me kaun:
- သဘင်ပညာ (tha- bin pa nja) ==> dha- bin pjin nja
- သဲကန္တရ (the: kan ta ja.) ==> the: gan da ja.
- အားကြီး (a: kyi:) ==> a: kyi:, a: gyi:

Corpora (myG2P)

19665	သုတိ	သူ တိ	thu. ti.
19666	သုတေသန	သူ တေ သ န	thu. tei tha- na.
19667	သုတေသီ	သူ တေ သ ီ	thu. tei thi
19668	သုဓမ္မဇရပ်	သူ ဓမ္မ မာ ဇ ရပ်	thu. da- ma za- ja'
19669	သုဓာဘုတ်	သူ ဓ ာ ဘုတ်	thou' da bou'
19670	သုနာပရန္တတိင်း	သူ နာ ပ ရန္တ တ တိင်း	thu. na pa- ran ta. tain:
19671	သုဘရာဇာ	သူ ဘ ရာ ဇ ာ	thu. ba. ja za
19672	သုမဂ်လ	သူ မဂ် ဂ လ	thu. min ga- la.
19673	သုမဂ်လဒုမဂ်လ	သူ မဂ် ဂ လ ဒ ု မဂ် ဂ လ	thu. min ga- la. du. min ga- la.
19674	သုဝဏ္ဏလိပ်	သူ ဝဏ္ဏ ဏ လိပ်	thu. wun na. lei'
19675	သုဝဏ္ဏသာမဇာတ်	သူ ဝဏ္ဏ ဏ သာ မ ဇ ာ တ ်	thu. wun na. tha ma. za'

Corpora (myG2P)

19663	သုတ	သု တ	thu. ta.	θ <u>u</u> t <u>ə</u>	
19664	သုတစာပေ		သု တ စ ပေ	thu. ta. sa pei	θ <u>u</u> t <u>ə</u> s <u>à</u> p <u>è</u>
19665	သုတိ	သု တိ	thu. ti.	θ <u>u</u> t <u>i</u>	
19666	သုတေသန		သု တေ သ န	thu. tei tha- na.	θ <u>u</u> t <u>è</u> θ <u>ə</u> n <u>ə</u>
19667	သုတေသီ	သု တေ သီ	thu. tei thi	θ <u>u</u> t <u>è</u> θ <u>i</u>	
19668	သုဓမ္မဇရပ်		သု ဓမ္မ မာ ဇ ရပ်	thu. da- ma za- ja'	θ <u>u</u> d <u>ə</u> m <u>à</u> z <u>ə</u> ja?
19669	သုဓဘုတ်		သု ဓ ဓ ဘုတ်	thou' da bou'	θ <u>u</u> u? d <u>à</u> bou?

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

Corpora (myG2P)

- myG2P က NLP Task အမျိုးမျိုးအတွက် အသုံးဝင်ပါတယ်
- Automatic Speech Recognition (ASR)
- Text to Speech (TTS)
- Text normalization, linguistics research, information retrieval etc.

- ၉၅-၁-၃၆၆၂၁၉ ==> ကိုးငါး တစ် သုံးခြောက်ခြောက်နှစ်တစ်ကိုး
- ၂၅၄၇၅၉၆၇။ ==> နှစ်ဆယ့်ငါးရက်မှာမွေး
- ၉၃၀ မှာတွေ့မယ် ==> ကိုးနာရီ ခဲ့ မှာတွေ့မယ်

Corpora (myG2P)

- ICCA-2015, Myanmar မှာ myG2P အတွက် ပထမဆုံး စာတမ်းကို ဖတ်ခဲ့
- Best Paper Award ကို ရရှိခဲ့
- မြန်မာဘာသာမေဒ ပညာရှင်တွေက ပြောကြတဲ့ အသံပြောင်းတဲ့ ပုံစတွေကို လွှဲလာခဲ့တယ်
- Exceptional rule တွေ အများကြီး ရှိနိုင်တာ ရှာဖွေသိရှိ
- ဥပမာ ရှင်ခုန် (jin khoun)

6.1. Pattern 1

If the vowel combination of first syllable is ‘ဃ/in/aun’ or ‘ဦ/in’ or ‘န/an/un’ or ‘မ/an/ein’ or ‘ဗ/e’ or ‘ဃ/e:’ or ‘ဃ/an’, and the consonant of second syllable is an unaspirated or an aspirated consonant that is unvoiced, then that second syllable’s pronunciation is voiced. Example pronunciations of some words are as follows:

တောင် ပုံး (taun pjoun:) => taun bjoun:
ပိုင်း ခြေ (pain: chei) => pain: gyei

Corpora (myG2P)

6.2. Pattern 2

If the first syllable is a stopped syllable (glottal stop) terminated with any of four vowel combinations (၁ါ, ၁ိါ, ၁၏ါ, ၁၂ါ), then the second syllable is pronounced independently from the first in the standard manner.

For example:

၁၅ ၁ီ: => se' thi:

The first syllable ၁၅ terminated with ၁ vowel combination, the second syllable ၁ီ: is pronounced, as it's standard pronunciation thi:

၁၅၁ ၁၁ => lhu' to

- The original “၁-t” does not change to voiced group.
- Approximately 50% of all cases are exceptions, for example the following case where the first syllable of a word ends with ၁/a', but the second syllable's pronunciation is changed to voiced.
- နတ်ကတော် (na' ka. to) => na' ga-do
- စုစုပေါင်း pattern ခုနစ်မျိုးအပေါ်လွှဲလာခဲ့တယ်

Corpora (myG2P)

- PACLING-2015, Indonesia မှာ myG2P ဒုတိယ စာတမ်းကို ဖတ်ခဲ့
- Conditional Random Fields (CRF) နဲ့ Phrase based Statistical Machine Translation (PBSMT) ကို သုံးခဲ့
- PBSMT ရဲ့ ရုလွှှိတွေက CRF ထက် ပိုကောင်းတယ်

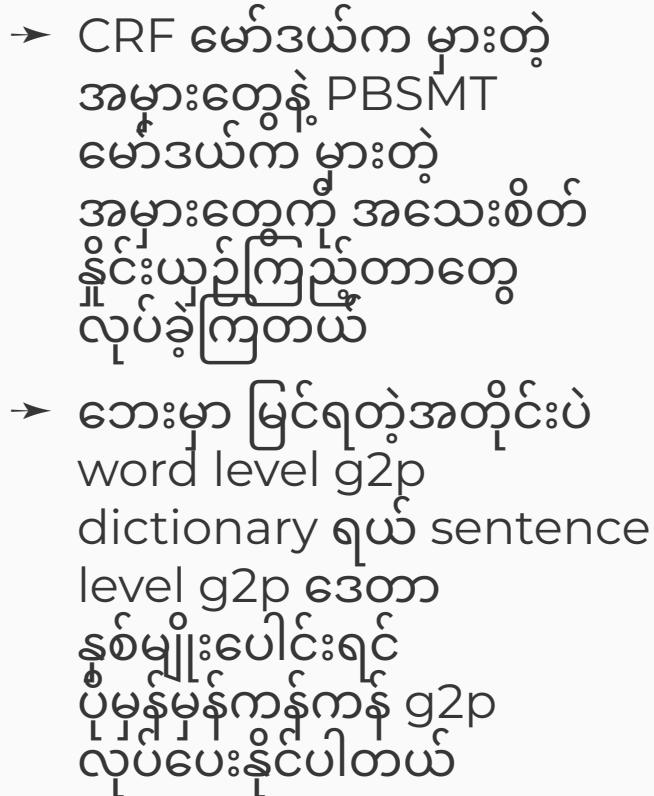
Table 5. Test set BLEU of Myanmar G2P

Test Data	Dictionary	Sentence	Dict + Sentence
Test Set 1	37.15	74.63	74.59
Test Set 2	34.70	73.73	74.18
Test Set 3	38.55	75.33	75.57
Test Set 4	80.34	79.17	86.29

Table 6. Phoneme accuracy of CRF and SMT approaches

Test-Data	CRF			SMT		
	dict	sentence	dict + sentence	dict	sentence	dict + sentence
Test Set 1	50.48	73.56	74.21	65.34	89.66	89.59
Test Set 2	49.60	73.82	74.36	63.64	89.24	89.45
Test Set 3	51.31	74.55	75.17	65.69	89.94	90.12
Test Set 4	75.93	72.71	77.71	92.79	91.85	94.29

Corpora (*myG2P*)

- CRF မော်ဒယ်က မှားတဲ့
အမှားတွေနဲ့ PBSMT
မော်ဒယ်က မှားတဲ့
အမှားတောက် အသေးစိတ်

လုပ်ခံကြတယ
- ဘေးမှ မြင်ရတဲ့အတိုင်းပဲ
word level g2p
dictionary ရယ် sentence
level g2p ဒေဝါး
နစ်မျိုးပေါင်းရင်
ပုံမှန်မှန်ကုန်ကန် g2p
လုပ်ပေးနှင့်ပါတယ်

ဒုက္ခ ခ ပဲ ဖြော တဲ့ စ ကား က မ တူ ဘူး ။

So bad, we are speaking in different languages.
dou' kha. be: pjo: de. za- ga: ga. ma- tu bu:

CRF (dictionary model):

dou' kha. **pe:** pjo: **te.** za- ga: **ka-** ma- tu bu:

CRF (sentence/dictionary+sentence model):

dou' kha. **be:** pjo: **de.** za- ga: **ga.** ma- tu bu:

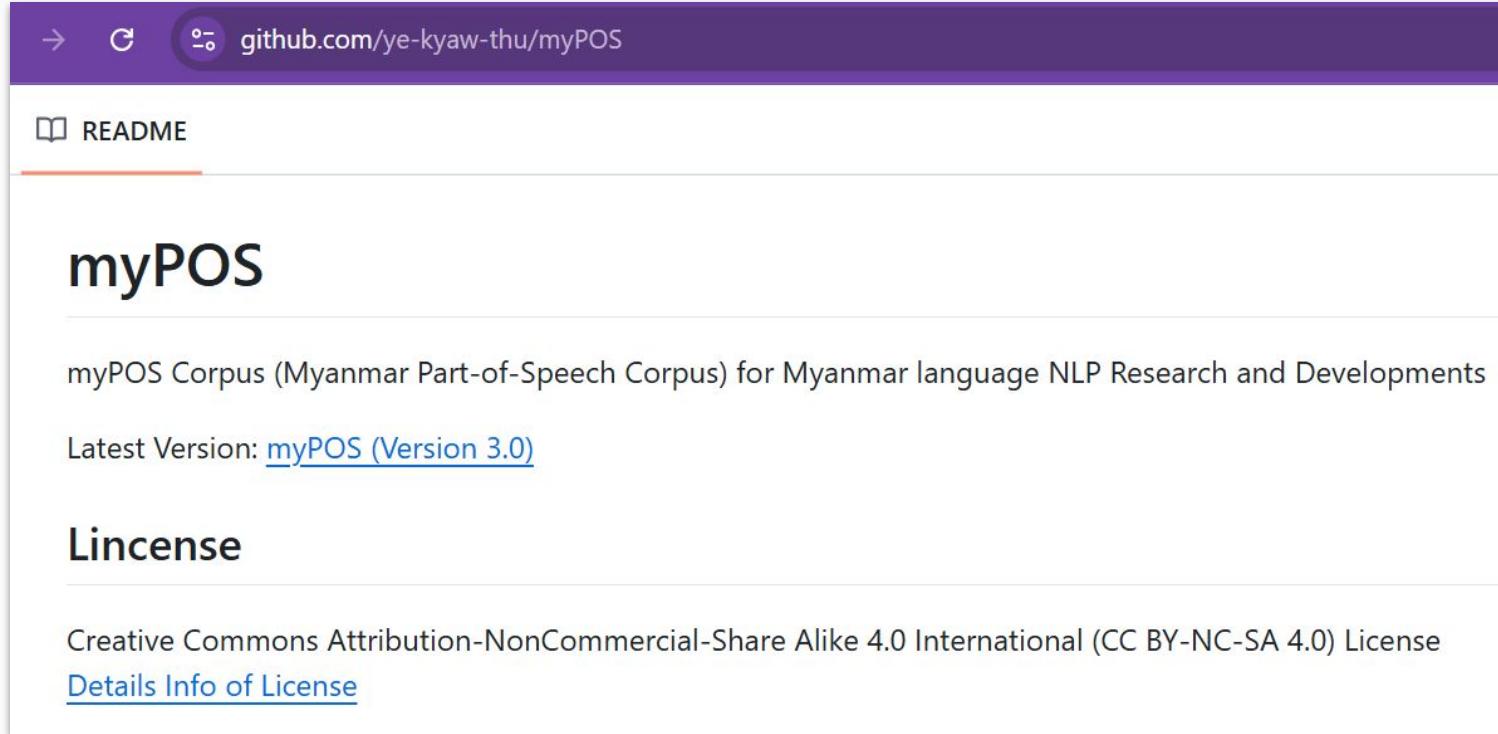
SMT (dictionary model):

dou' kha. **pe:** pjo: **te.** za- ga: **ka-** ma- **du** bu:

SMT (sentence/dictionary+sentence model):

dou' kha. **be:** pjo: **de.** za- ga: **ga.** ma- tu bu:

Corpora (myPOS)



A screenshot of a GitHub repository page for 'myPOS'. The page has a purple header with navigation icons and the URL 'github.com/ye-kyaw-thu/myPOS'. Below the header, there's a 'README' section with a link. The main content area features a large heading 'myPOS' and a description of the 'myPOS Corpus (Myanmar Part-of-Speech Corpus) for Myanmar language NLP Research and Developments'. It also mentions the 'Latest Version: [myPOS \(Version 3.0\)](#)'. A 'Lincense' section follows, stating the 'Creative Commons Attribution-NonCommercial-Share Alike 4.0 International (CC BY-NC-SA 4.0) License' and providing a link to '[Details Info of License](#)'.

→ github.com/ye-kyaw-thu/myPOS

README

myPOS

myPOS Corpus (Myanmar Part-of-Speech Corpus) for Myanmar language NLP Research and Developments

Latest Version: [myPOS \(Version 3.0\)](#)

Lincense

Creative Commons Attribution-NonCommercial-Share Alike 4.0 International (CC BY-NC-SA 4.0) License

[Details Info of License](#)

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

Corpora (myPOS)

- myPOS Corpus (Myanmar Part-of-Speech Corpus) for Myanmar language NLP Research and Developments.
- To the best of our knowledge, this corpus will be the biggest publicly available POS tagged dataset for Burmese or Myanmar language. We used 15 POS tags same as version 1.0 and 2.0. Please read the main README file (under myPOS/ folder) for the detail explanation of word segmentation and POS tags definition of the corpus. In this version 3.0, we extended the myPOS corpus of 11K sentences to 43,196 sentences by adding about 20K Myanmar sentences from our developing parallel corpora (i.e. Myanmar-Chinese and Myanmar-Korean) and 12K Myanmar sentences from ASEAN MT corpus of NECTEC, Thailand (See Table.1). All new raw sentences are manually segmented and POS tagging was done with myPOS RDR POS tagger. After that, we checked and fixed the POS tagging errors manually for the whole extended corpus.

Corpora (myPOS)

Table.1 Corpus information of myPOS (version 3.0)

Unit	myPOS (ver. 1.0)	Ext-1: my-zh	Ext-2: my-ko	Ext-3: ASEAN-MT my	myPOS (ver. 3.0)
Sentences	11,000	10,000	10,052	12,144	43,196
Words	239,598	103,909	106,864	114,134	564,505
Average Words/Sentence	21.78	10.17	10.64	9.40	13.07

- PhD ကျောင်းသူ နှစ်ယောက် ဖြစ်တဲ့ မခင်ဝါဝါထိက်၊ မအဏာလိုင် တိနှစ်ယောက်နဲ့အတူ ဒီ corpus ကို ဆောက်ဖြစ်ခဲ့တာပါ။ တခြား ကူညီပေးတဲ့ ဆရာမတွေလည်း ရှုပါတယ်။
- ပထမဆုံး ဗားရှင်းကို 2017 မှာ release လုပ်ဖြစ်ခဲ့တယ်။
- Version 3 ကိုတော့ ၂၀၂၀ မှာ release လုပ်ခဲုပါတယ်

Corpora (myPOS)

- ဒီ/adj ဆေး/n က/ppm ၁၀၀/num ရာခိုင်နှုန်း/n ဆေးဘက်ဝှံ/adj အပင်/n များ/part မှ/ppm ဖော်စပ်/v ထား/part တာ/part ဖြစ်/v တယ်/ppm ။/punc
- အသစ်/h စယ်/v ထား/part တဲ့/part ဆွဲယ်တာ/n က/ppm အသီး/n ထဲ/v နေ/part ပါ/part ပေါ့/part ။/punc
- မ/part ကျိန်းမာ/v လျှင်/conj နတ်/h|ဆရာ/h ထံ/ppm မေးမြန်း/v ၍/conj သက်ဆုံးရာ/h နတ်/h တို့/part အား/ppm ပူဇော်ပသ/v ရဲ/part သည်/ppm ။/punc
- ပေဟုံဗ္ဗား/n|ဦးယူရှုံး/n ။/punc
- နိုမ်/adj အပ်မက/h ကောသလ/h|မင်း/h|အိပ်မက်/h ၉/num နှက်ရိုင်း/adj ကျယ်ဝန်း/adj သော/part ရေကန်/h ကြီး/adj တစ်/tn ခု/part တွင်/ppm သတ္တဝါ/h တို့/part ဆင်း/v ၍/conj ရေသောက်/v ကြား/part ၏/ppm ။/punc

Corpora (myPOS)

Methods	Closed Test-set	Open Test-set
CRFs	97.77%	95.05%
HMM	97.31%	96.43%
MaxEnt	96.55%	96.31%
RDR	98.42%	<u>97.05%</u>
SVM	<u>99.83%</u>	93.55%
Two-Hours	95.83%	92.87%

Fig. POS tagging result of six methodologies (with myPOS corpus ver 0.9)

Corpora (myPOS)

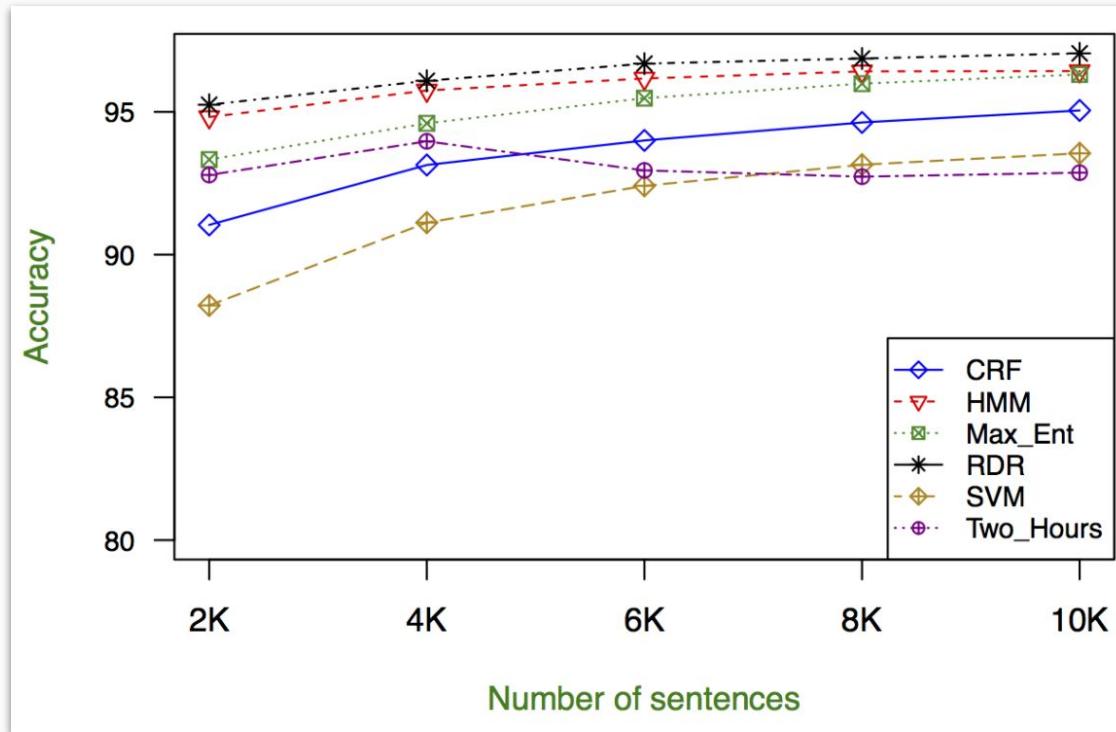


Fig. Accuracies of six POS tagging methodologies on varying training data sizes

Corpora (myPOS)

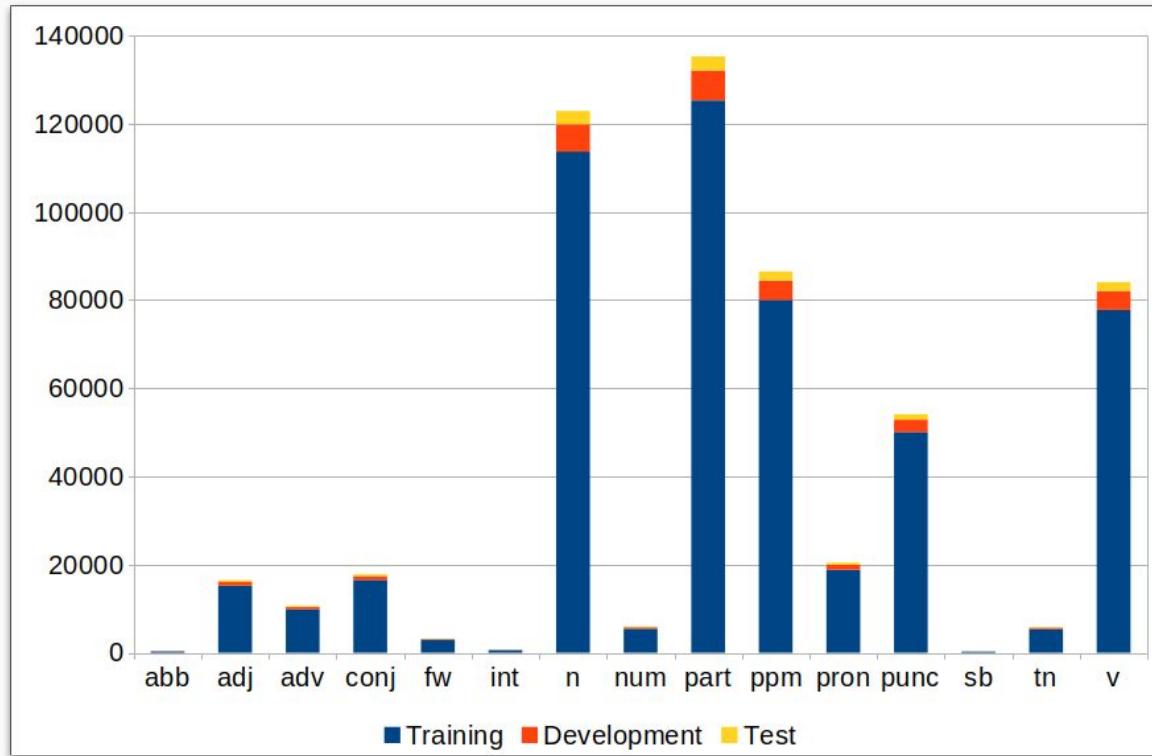


Fig. POS tag distribution of myPOS corpus version 3.0

Corpora (myPOS)

- Khin War War Htike, Ye Kyaw Thu, Zuping Zhang, Win Pa Pa, Yoshinori Sagisaka and Naoto Iwahashi, "Comparison of Six POS Tagging Methods on 10K Sentences Myanmar Language (Burmese) POS Tagged Corpus", at 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), April 17~23, 2017, Budapest, Hungary.
- Zar Zar Hlaing, Ye Kyaw Thu, Myat Myo Nwe Wai, Thepchai Supnithi, Ponrudee Netisopakul, "Myanmar POS resource extension effects on automatic tagging methods", In Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2020), Nov 18 to Nov 20, 2020, Bangkok, Thailand, pp. 189-194.

Corpora (myNER)

The screenshot shows a GitHub repository page for 'myNER' at github.com/ye-kyaw-thu/myNER. The repository has two files: '7-tags' and 'README.md'. Both were updated 4 months ago. The 'README' file is currently selected, showing its content.

myNER

Named Entity Recognition (NER) corpus for Burmese (Myanmar language)

The Language Understanding Lab is developing NER corpora with both 7-tag and 23-tag annotation schemes. Currently, we have released only the [myNER \(7-tags\)](#) corpus.

→ <https://github.com/ye-kyaw-thu/myNER>

Corpora (myNER)

Table 1. Tag Names, Descriptions, and Examples of myNER Corpus

Tag	Description	Examples
DATE	Date	၁ ရက် ၅ လ ၁၉၉၆ (01-05-1996), ကဆုန်လပြည့် (full moon day of Kasone)
TIME	Time	မနက် ၆ နာရီ ၁၀ မိနစ် (06:10 AM)
NUM	Numbers	၁ (1), ၁သိန်း (1 Million)
PER	Person Names	အန်တို့နယို (Antonio), ကောင်းလွင်သန့် (Kaung Lwin Thant)
ORG	Organizations and Institutions	မူက်ခရီးဆော့ (Microsoft Inc), ဟဘံဗတ်တက္ကသိုလ် (Harvard University)
LOC	Locations and Geographic Features	အာရှတိုက် (Asia), ထိုင်းနိုင်း (Thailand)
O	Outside	Nonspecific entities which do not match the above tags

- myNER corpus containing 16, 605 sentences, originally sourced from data taken from the myPOS version 3 corpus
- လောလောဆယ်မှာတော့ 7-tag NER ကိုပဲ release လုပ်ထားပါတယ်

Corpora (myNER)

```
261 ဒီမယ်>int>O  
262 |>punc>O  
263 ရန်ကုန် n>B-LOC  
264 ပြည့်သူ့ n>I-LOC  
265 ဆေးရုံကြီး>n>E-LOC  
266 ဘယ်>pron>O  
267 မှ>ppm>O  
268 လဲ>part>O  
269 ||>punc>O  
270  
271  
272 ရွာ>n>O  
273 နေရာ>n>O  
274 ကုတ်>n>O  
275 မှ>ppm>O  
276 ၁၇၁၃၀->num>S-NUM  
277 ဖြစ်>v>O  
278 သည်>ppm>O  
279 ||>punc>O  
280  
281 မိသုဇာဝ >n>S-PER
```

- ငြော့ format က Conference on Natural Language Learning (CoNLL) ပါ
- ကော်လံ သုံးခု သုံးထားတယ်
- ပထမဆုံး ကော်လံမှာက စာလုံးဖြတ်ထားတဲ့ မြန်မာစာ စာလုံး
- ဒုတိယ ကော်လံမှာက Part-of-Speech (POS) tag
- တတိယ ကော်လံမှာက NER tag
- စာကြောင်း တစ်ကြောင်းနဲ့ တစ်ကြောင်းအကြားမှာ blank line ခြား

Corpora (myNER)

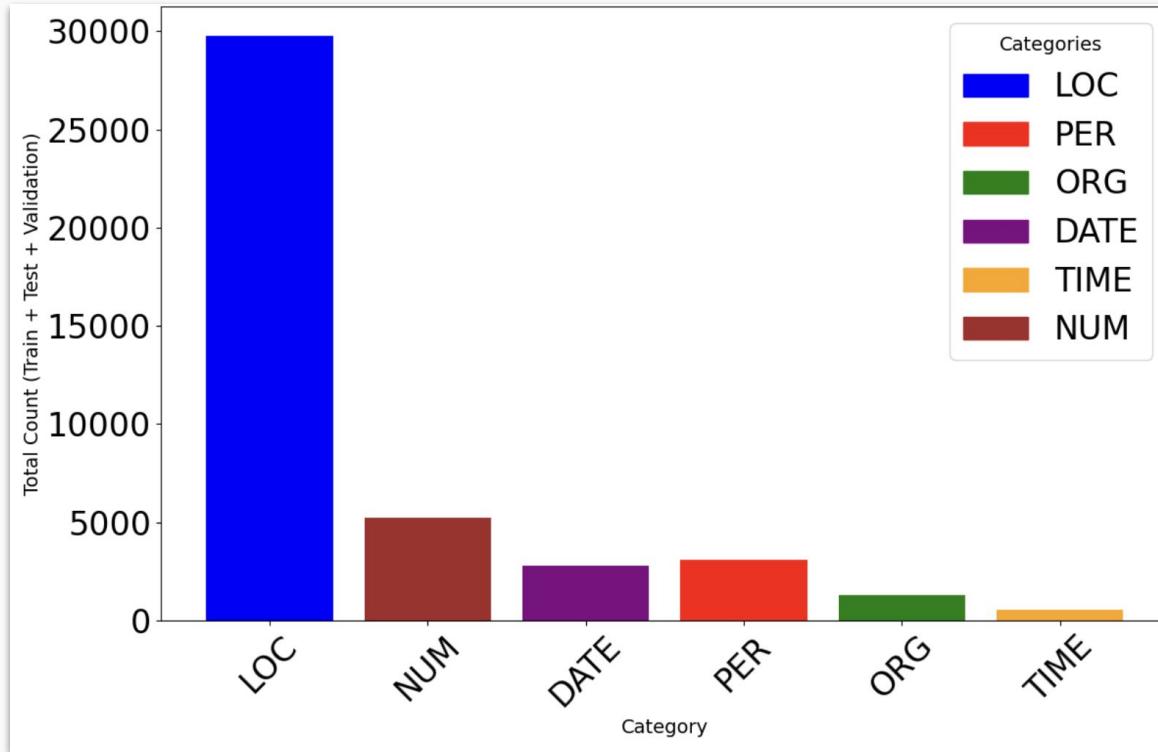
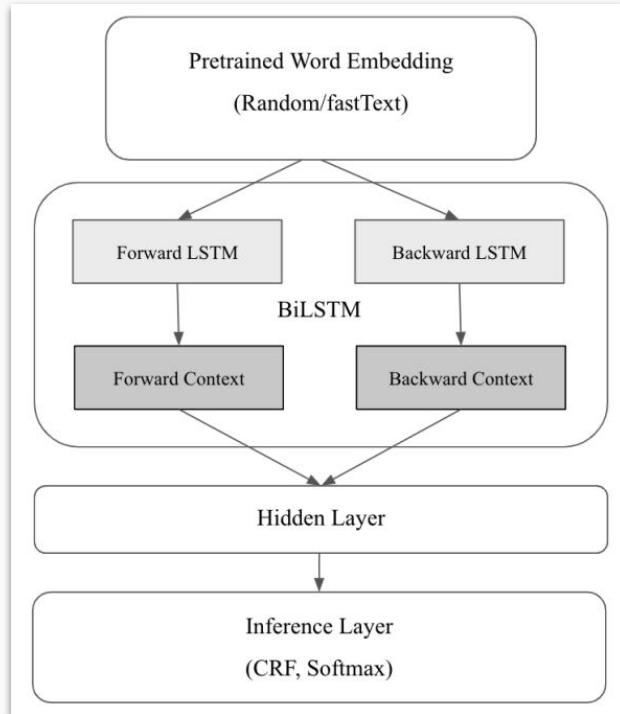


Fig.Distribution of NER tag categories in the myNER corpus (excluding the 'O' tag).

Corpora (myNER)



- မော်ဒယ်အနေနဲ့က Conditional Random Fields (CRF) နဲ့ Bidirectional LSTM (BiLSTM)-CRF ကို သုံးခဲ့တယ်
- fastText embeddings ကိုလည်း သုံးခဲ့တယ်
- NER tagging တစ်ခွဲတည်း မဟုတ်ပဲ POS tagging ကိုပါ တွဲပြေး လုပ်တဲ့ နည်းလမ်းကို လေ့လာခဲ့တယ်
- ဆိုလိုတာက POS information ကိုလည်း မြှောင်းနဲ့ NER လုပ်သွားတဲ့ ပုံစံမျိုးပါ

Fig.Word Embeddings + BiLSTM Sequence Tagging

Corpora (myNER)

Model Name	Embeddings	Training Type	Acc.	F1-Wt.	F1-Macro.
CRF	w/o fastText features	Single (NER)	0.9818	0.9812	0.7405
		Joint (POS+NER)	0.9812	0.9807	0.7367
	with fastText features	Single (NER)	0.9818	0.9811	0.7429
		Joint (POS+NER)	0.9810	0.9804	0.7345
BiLSTM-Softmax	+ Random Embeddings	Single (NER)	0.9740	0.9725	0.6478
		Joint (POS+NER)	0.9730	0.9714	0.6463
	+ frozen fastText	Single (NER)	0.9737	0.9723	0.6578
		Joint (POS+NER)	0.9753	0.9734	0.6489
	+ fine-tuned fastText	Single (NER)	0.9783	0.9779	0.6502
		Joint (POS+NER)	0.9780	0.9764	0.6743
BiLSTM-CRF	+ Random Embeddings	Single (NER)	0.9740	0.9730	0.6784
		Joint (POS+NER)	0.9742	0.9730	0.6907
	+ frozen fastText	Single (NER)	0.9747	0.9729	0.6396
		Joint (POS+NER)	0.9746	0.9737	0.6790
	+ fine-tuned fastText	Single (NER)	0.9755	0.9753	0.7154
		Joint (POS+NER)	0.9791	0.9776	0.7395

Table. Performance Comparison of Single NER and Joint POS+NER Models (Best results are bolded.)

→ နှစ်မျိုးတဲ့ပေမူ
ရလဒ်က
အတူတူလောက်ပဲ
ရ

→ fine-tuned
fastText
embedding ကို
ပြုအရှင်စိုက်ခဲ့

→ Linguistic
feature ဖြည့်ချင်

Corpora (myNER)

- myNER corpus တည်ဆောက်ပြီး ဘယ်လောက် မှန်မှန်ကုန်ကုန် tag လုပ်ပေးနိုင်သလဲ ဆိုတာနဲ့ ပတ်သက်ပြီး အသေးစိတ် သချင်ရှင် အောက်ပါ စာတမ်းကို ဖြင့်မြောက်ပါ။
- 28-Nov-2023: Our workshop paper "myNER9: Development, Manual Annotation, and Evaluation of a 9-Tag Myanmar NER Corpus via XGBoost and Bi-LSTM" achieved "**the Best Paper for Workshop**" at the 5th NLP/AI R&D Workshop, 2023, Bangkok, Thailand.
- Kaung Lwin Thant, Kwankamol Nongpong, Ye Kyaw Thu, Thura Aung, Khaing Hsu Wai, Thazin Myint Oo , "myNER: Contextualized Burmese Named Entity Recognition with Bidirectional LSTM and fastText Embeddings via Joint Training with POS Tagging ", the International Conference on Cybernetics and Innovations (ICCI 2025), April 2-4, Pattaya Chonburi, Thailand pp.1-6 (**Best Presentation Award**)

Corpora (myUDTree)

myUDTree

Introduction

The myUDTree corpus is a Universal Dependency (UD) Corpus that extends previous work of the Myanmar UD parsing (Hnin Thu Zar Aye et al., 2018), including 11,000 sentences of dependency tree data. The extended myUDTree corpus contains 43,196 sentences in total.

Universal Dependencies (UD) Corpus is a type of corpus that is annotated according to the grammar rule of the respective languages with Part-of-Speech (POS), morphological features, and syntactic dependencies. Before building myUDTree, we built a [Joint POS Tagging and Graph-based Dependency Parsing \(jPTDP\)](#) model, using the existing in-house version of Myanmar UD corpus in order to parse raw data with dependency information. After building the jPTDP model, we selected 20,000 Myanmar sentences from our developing parallel corpora (i.e., from Myanmar-Chinese and Myanmar-Korean language pairs), and 12,000 Myanmar sentences from the ASEN-MT corpus built by [NECTEC Research Center](#) in Thailand, and these data are parsed by using our built jPTDP model to obtain dependency-tree data.

→ <https://github.com/ye-kyaw-thu/myUDTree>

Corpora (myUDTree)

Universal Dependency (UD) corpus ဆိုသည်မှာ Part-of-Speech (POS) အပြင် morphological feature များ၊ syntactic dependency များဖြင့် သက်ဆိုင်ရာ ဘာသာစကားရဲ့ grammar rule အရ annotated လုပ်ထားတဲ့ corpus အမျိုးအစားဖြစ်ပါတယ်။ myUDTree ကို မတည်ဆောက်ခင်မှာ dependency information များဖြင့် raw data များကို parsing လုပ်နိုင်ရန်အတွက် မူလရှိပြီးသား စာကြောင်းရေး တစ်သောင်းကျော်သာ ရှိတဲ့ Myanmar UD corpus အသုံးပြု၍ dependency parsing model တစ်ခုဖြစ်တဲ့ [Joint POS Tagging and Graph-based Dependency Parsing \(jPTDP\)](#) မော်ဒယ် ကို တည်ဆောက်ခဲ့ပါတယ်။

jPTDP မော်ဒယ်ကို တည်ဆောက်ပြီးတဲ့ နောက် လက်ရှိမှာ machine translation သုတေသနအတွက် ပြင်ဆင်နေတဲ့ parallel corpora (i.e. Myanmar-Chinese and Myanmar-Korean) မှ Myanmar raw data စာကြောင်းရေး ၂၀,၀၀၀ နှင့် ထိုင်းနှင့် [NECTEC Research Center](#) မှ တည်ဆောက်ထားသော ASEAN MT Corpus မှ Myanmar raw data စာကြောင်းရေး ၁၂,၀၀၀ တို့အား jPTDP model ဖြင့် dependency tree data များရရှိရန် parsing လုပ်ခဲ့ပါတယ်။ Parsed လုပ်ထားသော ဒေတာများကို CoNLL-U Viewer tool ကို အသုံးပြုပြီး manual အားဖြင့် အချိန်ယူ စစ်ဆေးပြီးပြင်ဆင်ထားပါတယ်။ အဲဒေတာများရရှိရန်မှာ စစ်ဆေးပြင်ဆင်ထားသော dependency tree data များနှင့် မူလရှိပြီးသား Myanmar UD Corpus ကို parse လုပ်ထားတဲ့ output စာကြောင်းတွေနဲ့ပေါင်းလိုက်ပြီး စာကြောင်းရေး အရေအတွက် ၄၃,၁၆၉ နှင့် dependency tree information များပါဝင်သော myUDTree Corpus (version 1.0) ကို ဖွံ့ဖည်းထားပါတယ်။

Corpora (myUDTree)

Myanmar dependency structure မှ အများဆုံး အသုံးပြုတဲ့ dependency relation တွေကတော့
အောက်ဖော်ပြပါ relation များဖြစ်ကြပါတယ်။

- root (root)
- acl (clausal modifier of noun)
- amod (adjectival modifier)
- advmod (adverbial modifier)
- case (case marking)
- mark (marker)
- compound (compound)
- obl (oblique nominal)
- obj (object)
- punct (punctuation)

Corpora (myUDTree)

Table 1. Corpus information of the myUDTree (version 1.0)

Unit	myPOS (ver. 1.0)	Ext-1: my-zh	Ext-2: my-ko	Ext-3: ASEAN-MT (my)	myUDTree (ver. 1.0)
Sentences	11,000	10,000	10,052	12,144	43,196
Words	239,598	103,909	106,864	114,134	564,505
Average Words/Sentence	21.78	10.17	10.64	9.40	13.07

- myUDTree (Myanmar Universal Dependency Tree) ၆၃၀၁၂၅ ၂၀၂၂ ခုနစ်မှာ အများသုံးလိုရအောင် release လုပ်ခဲ့ပါတယ်။

Corpora (myUDTree)

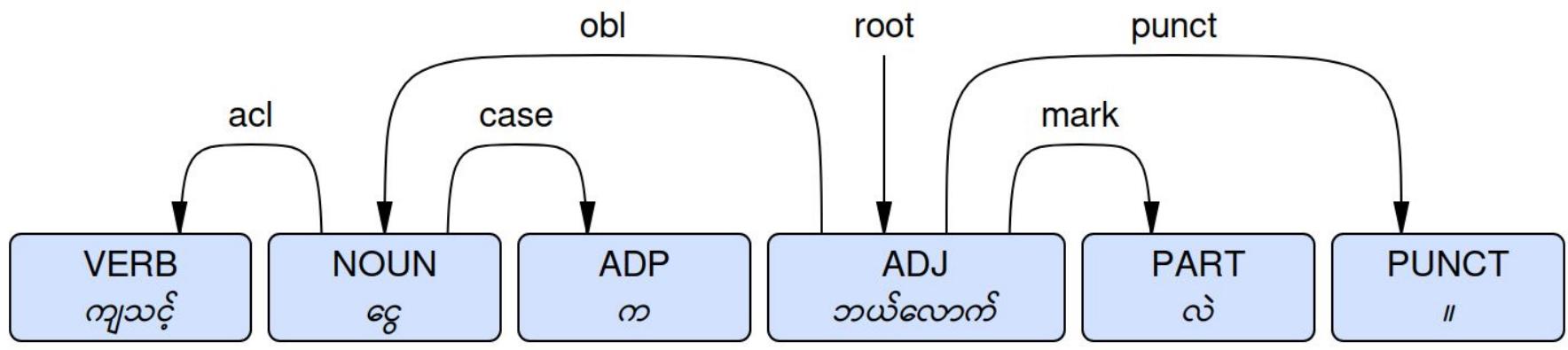
Example sentence no.1

Input:

1	ကျသင့်	-	VERB	-	-	2	acl	-	-
2	ငွေ	-	NOUN	-	-	4	obl	-	-
3	က	-	ADP	-	-	2	case	-	-
4	ဘယ်လောက်	-	ADJ	-	-	0	root	-	-
5	လဲ	-	PART	-	-	4	mark	-	-
6	။	-	PUNCT	-	-	4	punct	-	-

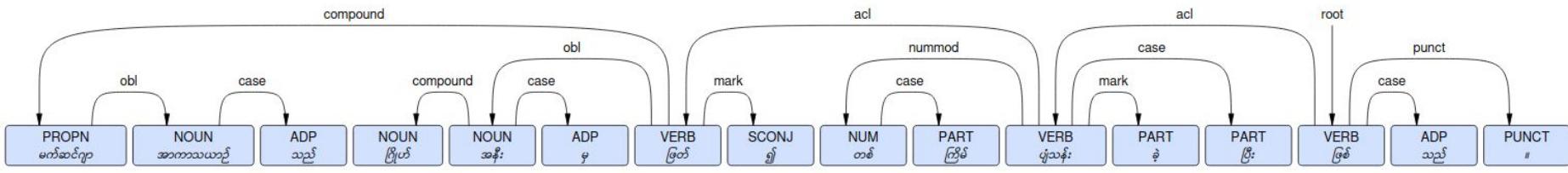
→ CoNLL-U Viewer ကို အထက်ပါ format နဲ့ input ပေးလိုက်ရင်

Corpora (myUDTree)



→ അംഗീകാരത്തിൽ: dependency tree ബുന്ധി തൃത്യപേഖ്യിലിൽമുണ്ട്

Corpora (myUDTree)



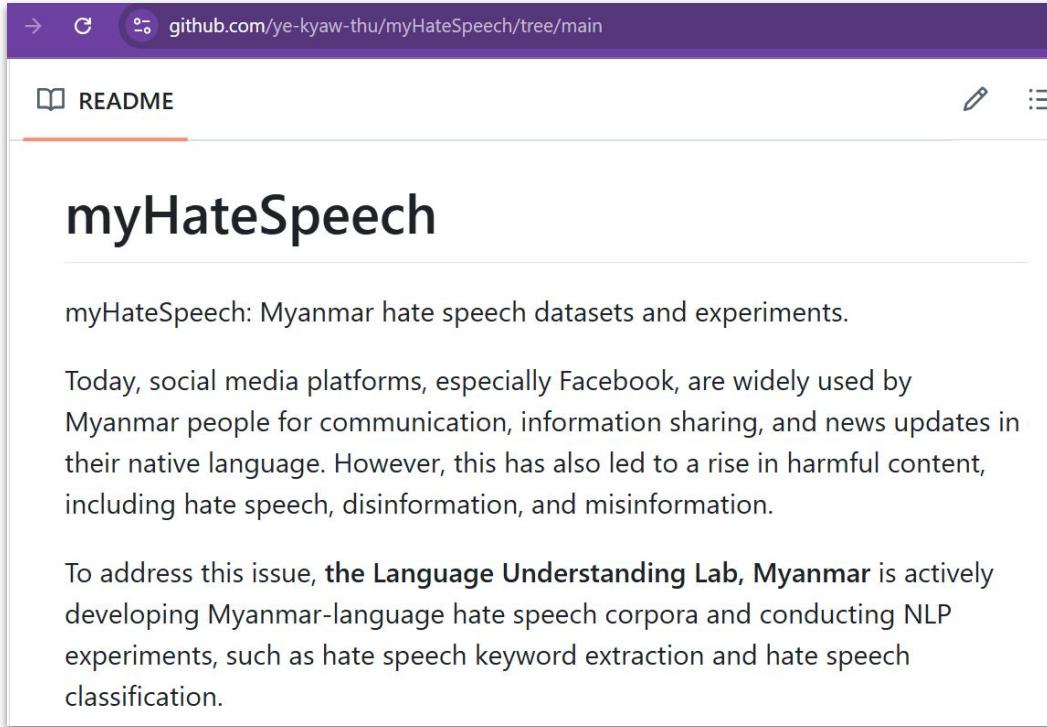
1	မက်ခင်္ဂာ	-	PROPN	-	-	7	compound	-	-
2	အာကာသယှဉ်	-	NOUN	-	-	1	obl	-	-
3	သည်	-	ADP	-	-	2	case	-	-
4	မြတ်	-	NOUN	-	-	5	compound	-	-
5	အနီး	-	NOUN	-	-	7	obl	-	-
6	မ	-	ADP	-	-	5	case	-	-
7	ဖြတ်	-	VERB	-	-	11	acl	-	-
8	၏	-	SCONJ	-	-	7	mark	-	-
9	တစ်	-	NUM	-	-	11	nummod	-	-
10	ကိမ်	-	PART	-	-	9	case	-	-
11	ပုံသန်း	-	VERB	-	-	14	acl	-	-
12	ခဲ့	-	PART	-	-	11	mark	-	-
13	ပြီး	-	PART	-	-	11	case	-	-
14	ဖြစ်	-	VERB	-	-	0	root	-	-
15	သည်	-	ADP	-	-	14	case	-	-
16	။	-	PUNCT	-	-	14	punct	-	-

- ပုံမှာ မြင်ရတဲ့အတိုင်းပါပဲ စာကြောင်းက ရှိလျှောတာနဲ့အမျှ dependency tree ရဲ့ တည်ဆောက်ပံ့ကလည်း ပိမ့်ရတ်ထွေးနဲ့ဝါတာ၊ ပုံစံ အမျိုးမျိုး ရှိတာ ဖြစ်နှင့်ပါတယ
- တကယ်ကို ပင်ပင်ပန်ပန်း အာရုံစိုက်ပြီး ဒီ ကောပတ်စံကို ဆောက်ခဲ့ကြပါတယ
- ဒါပေမဲ့ မှားနေတာတွေ ရှိနှင့်ပါသေးတယ်

Corpora (myUDTree)

- myUDTree (version 1.0) နဲ့ပတ်သက်တဲ့ စာတမ်းက အောက်ပါအတိုင်းပါ
- Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi and Ponrudee Netisopakul, "Graph-based Dependency Parser Building for Myanmar Language", In Proceedings of the 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2022), Nov 5 to 7, 2022, Chiang Mai, Thailand, pp. 1-6.
- Creative Commons Attribution-NonCommercial-Share Alike 4.0 International (CC BY-NC-SA 4.0) License
- လိုင်စင်နဲ့ပတ်သက်တဲ့အသေးစိတ်က ဒီလင့်ကို ဖတ်ပါ။
<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Corpora (myHateSpeech)



The screenshot shows the GitHub repository page for 'myHateSpeech'. The title 'myHateSpeech' is displayed prominently. Below it, a brief description states: 'myHateSpeech: Myanmar hate speech datasets and experiments.' A paragraph explains that social media platforms like Facebook are used by Myanmar people for communication but also note a rise in harmful content such as hate speech. Another paragraph discusses the work of the Language Understanding Lab, Myanmar, which is developing hate speech corpora and conducting NLP experiments.

→ <https://github.com/ye-kyaw-thu/myHateSpeech>

- အထူးသဖြင့် social media တွေထက
မြန်မာအများစု
အသံးပြုကြတဲ့ Facebook
မှာ ဆုရင် နေစဉ်လိုလို hate
speech တွေကဲ ပုစ
အမျိုးမျိုးနဲ့
တွေကြရပါလိမ့်မယ်
- အဲဒါတွေကို ကုန်ပျူတာနဲ့
automatic detect
လုပ်တာမျိုး လုပ်နိုင်ဖို့
ရည်ရွယ်တယ်

Corpora (myHateSpeech)

- ဖော်လော်မော်/ab မ ဟုတ် လို့ ပေါ့ 😂 😂 ab
- နား ကို မ လည် တာ no
- ဆောက်မြင်ကပ်/ab ထင် တာ ပဲ ab
- ကွွမ်းယာ မှာ ထည့် စား တဲ့ စမုန်စပါး ထင် တယ် no
- ဖလော်မော်/ab next version ab
- ငါ လည်း သိ ချင် နေ တာ 😊 အဲလို့ စကား တွေ ကျ နားမလည် လို့ သင် ပေး ကြ ပါ ဦး 😂 no
- ငါ မ သိ လို့ ကိုကို ကို မေး ကြည့် တာ ကိုကို က လည်း baby က လွှဲ ရင် မ သိ ဘူး တဲ့ 😔 no
- ဖော်လော်မော်/ab နဲ့ ညီမ တော် တယ် လေ 😊 ab
- ဒမ္မား/ab ကြီး တဲ့ 😂 ab
- \$မ္မား/ab ပါ ab

Corpora (myHateSpeech)

- myHateSpeech ကောပတ်စက်လည်း ဒေတာဆက်တိုးဖို့ လိုအပ်ပါတယ်။ ဆက်လည်း လုပ်သွားပါမယ်။
- လောလောဆယ် ဆောက်ပြီးသလောက် ဒေတာကိုပဲ သုံးပြီး လုပ်ခဲ့တဲ့ စမ်းသပ်မှုရလဒ်တွေကိုအာက်ပါ စာတမ်းအနေနဲ့ ရေးသားထုတ်ဝေခဲ့ပါတယ်။
- Nang Aeindray Kyaw, Ye Kyaw Thu, Hutchatai Chanlekha, Thazin Myint Oo, Okumura Manabu and Thepchai Supnithi, "Enhancing Hate Speech Classification in Myanmar Language Through Lexicon-Based Filtering", the 21st International Joint Conference on Computer Science and Software Engineering (JCSSE 2024), June 19-22, Phuket, Thailand, pp. 325-332
- myHateSpeech ကောပတ်စက် စိတ်ဝင်စားတဲ့သူတွေ၊ ကူချင်တဲ့ သူတွေ ရှိရင်လည်း ကျွန်တော့ကို ဆက်သွယ်ပါ။

Corpora (myContradict)



The screenshot shows a GitHub repository page for 'myContradict'. The URL in the address bar is github.com/ye-kyaw-thu/myContradict. The page title is 'myContradict: Contradictory Sentence Generation for Myanmar language'. Below the title, there's a section titled 'Overview' with the following text:

This dataset contains pairs of sentences in the Myanmar language, where each pair consists of an original sentence and its manually created contradictory version. The dataset is designed to support research in natural language understanding, particularly in the area of sentence contradiction and semantic analysis. The source files are noted with src and target files are with tgt for both syllable- and word-level parallel corpora. We also put POS-Tagged corpora, which was used in our experiments. For the experiments, we used OpenNMT and we shared our yaml files for baseline experiments.

→ <https://github.com/ye-kyaw-thu/myContradict>

→ Conversational AI
အလုပ်တွေအတွက်
တဖက်က ပြောလိုက်တဲ့
စကားကို ဆန့်ကျင့်ပြီး
ပြောနိုင်တာကလည်း
အရေးကြီးလို့
myContradict ဒေတာကို
ပြင်ဆင်ဖြစ်ခဲ့တယ်

Corpora (myContradict)

Dataset Statistics

Split	No. of Syllables	No. of Words	No. of Sentences
Train	150,986 (Source) / 151,007 (Target)	112,917 (Source) / 112,900 (Target)	9,702 (Source) / 9,702 (Target)
Valid	18,613 (Source) / 18,622 (Target)	14,050 (Source) / 14,043 (Target)	1,214 (Source) / 1,214 (Target)
Test	19,418 (Source) / 19,392 (Target)	14,498 (Source) / 14,499 (Target)	1,214 (Source) / 1,214 (Target)

- myContradict ဒေတာမှာ ပါဝင်တဲ့ စုစုပေါင်း ဝဏ္ဏ၊ စာလုံး နဲ့ စာကြောင်း အရေအတွက်

Corpora (myContradict)

Segmentation Examples

The dataset includes different segmentations of sentences, both with and without Part-of-Speech (POS) tags. Below is an example of the word "ကရှုဏာ သက်" (compassionate) segmented in various ways:

Description	Segmentation Example
Syllables without POS Tags	က ရှ ဏာ သက်
Syllables with POS Tags	က B-n ရ I-n ဏ E-n ာ S-n သ S-v က် V
Words without POS Tags	ကရှုဏာ သက်
Words with POS Tags	ကရှုဏာ N သက် V

- myContradict ဒေတာ ရဲ့ word segmentation, tagging ဥပမာပါ။

Corpora (myContradict)

1. Negation:

- Example: **ql|v õ|part τay|ppm** → **ω|part ql|v õ|v ɔ̄l|ppm**
(Translation: *It is okay.* → *It is not okay.*)
 - In negation, the verb is modified with a negative particle (**ω**), and post-positional markers are adjusted.

2. Antonyms:

- Example: ○ v တဲ့ | part ပုဂ္ဂၢ၍ | n က | ppm ကျိုးမ်း | pron တို့ | part အငောင် | n ပါ | ppm → ကျိုးမ်း | pron တို့ | part အငောင် က | part စိန် | v ပါ | part တယ် | ppm
(Translation: *Fat person is our father.* → *Our father is thin.*)
 - Here, the verb ○ (fat) is replaced with its antonym စိန် (thin).

3. Counterpart:

- Example: မန္တလေး|n သို့|ppm သွား|v သည်|ppm → မန္တလေး|n မှ|ppm လာ|v သည်|ppm
(Translation: *He goes to Mandalay.* → *He comes from Mandalay.*)
 - This involves changing post-positional markers and verbs to indicate temporal or directional opposites.

Corpora (myContradict)

POS Tagging

- For training multi-feature models, POS tagging was performed using the myPOS version 3.0 model, which leverages the RDRPOSTagger and achieves an F1-score of 96.53%. POS tags are provided at the word level, and for syllable-level tagging, a BIES (Begin, Inside, End, Single) scheme is used.
- ဒီမှာပြောနေတဲ့ RDR ဆိုတာက Ripple Down Rules ဆိုတဲ့ နည်းလမ်းပါ

Corpora (myContradict)

126 ကျွန်မ မ ရွှေ့ဖျင်း ပါ ဘူး။ ကျွန်မ မ ထက်မြက် ပါ ဘူး။
 127 ဘယ် အချိန် ထိ မ ရ နိုင် ဘူး လဲ။ → ဘယ် အချိန် မှာ ရ နိုင် မလဲ။
 128 ဟုတ် ပါ တယ်။ ဟို ဘက် ကို ကြွာ ပါ ။ → ဟုတ် ပါ တယ်။ ဒီ ဘက် ကို ကြွာ ပါ ။
 129 သူ အရေး မ ဝ ဘူး။ → သူ သိပ် ဝ တယ်။
 130 ဂိတ်စ် တို့ ၏ အစမ်း သရုပ်ပြု မူ သည် အောင်မြင် ခဲ့ သည်။ → ဂိတ်စ် တို့ ၏ အစမ်း သရုပ်ပြု မူ သည် မ အောင်မြင် ခဲ့ ပါ။
 131 ခင်ဗျား တို့ ရဲ့ စက်ရုံ ကို သွား ကြည့် လို့ ရ သလား။ → ခင်ဗျား တို့ ရဲ့ စက်ရုံ ကို သွား မ ကြည့် လို့ ရ သလား။
 132 ကိုယ့် ရဲ့ ခံယူ ချက် ကို ပြင်ဆင် သင့် လည်း ခေါင်းမာ တယ်။ ကိုယ့် ရဲ့ ခံယူ ချက် ကို ပြင်ဆင် သင့် ရင် ပြင် ရ တယ်။
 133 ကြုက်ရုံ ကို လျှော့ ခလောက် ပါ ။ ကြုက်ရုံ ကို နာနာ ခလောက် ပါ ။
 134 ရ ပါ တယ် ။ လာ မှာ ပေါ့။ → ရ ပါ တယ် ။ မ လာ ပါ ဘူး။
 135 မိုက်မဲ လု ချုပ် လား။ → ကောင်း လု ချုပ် လား။
 136 ရေးဟောင်း နှင့်တော် ပြုတိုက် မှာ ပြုခန်း အမျိုးမျိုး ခဲ့ ထား ပါ တယ်။ → ရေးဟောင်း နှင့်တော် ပြုတိုက် မှာ ပြုခန်း တစ် ခု ပဲ ရှိ တယ်။
 137 ရေဒီယို နဲ့ ပြည့်တွင်းပြည့်ပုံ သတင်း တွေ အားလုံး ကို တော့ နားထောင် လို့ မ ရ ဘူး။ → ရေဒီယို နဲ့ ပြည့်တွင်းပြည့်ပုံ သတင်း တွေ အားလုံး ကို နားထောင် လို့ ရ တယ်။
 138 ကမြင်းကြောထ နေ ပြန် ပြီ။ → ငြိမ်ငြိမ်သက်သက် နေ ပြန် ပြီ။
 139 အဆုံးသတ် လိုက် ရအောင် ကျေးဇူးပြု၍ ကော့တေး အရက် မ ပေး ပါ နဲ့ တော့။ → စ လိုက် ရအောင် ကျေးဇူးပြု၍ ကော့တေး အရက် ပေး ပါ ။
 140 ကျွန်တော် တို့ အဲဒီ မှာ ပဲ ကျင်းပဲ ပါ မယ်။ → ကျွန်တော် တို့ အဲဒီ မှာ မ ကျင်းပဲ တော့ ပါ ဘူး။
 141 ကျွန်မ က တော့ ကလေး ရဲ့ အဖြေါး ကို နှစ်သက် တယ်။ → ကျွန်မ က တော့ ကလေး ရဲ့ အဖြေါး ကို မ ကြိုက် ဘူး။

→ စာကြောင်းနှစ်ကြောင်းအကြား TAB ခြားထားတဲ့ format ပါ

Corpora (myContradict)

- myContradict (version 1.0) ကောပတ်စ်ကိုသုံးပြီး လုပ်ခဲ့တဲ့ စမ်းသပ်မှု ရလဒ်တွေကိုတော့ အောက်ပါ စာတမ်းမှာ အသေးစိတ် ဖော်ပြထားပါတယ်
- Ye Kyaw Thu, Ei Myat Nwe and Thura Aung,
"myContradict: Semi-supervised Contradictory Sentence Generation for Myanmar language", In Proceedings of the 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2024), Nov 11 to 15, 2024, Pattaya, Thailand, pp. 1-6

Corpora (myParaphrase)

The screenshot shows a GitHub repository page for 'myParaphrase'. The header includes the URL 'github.com/ye-kyaw-thu/myParaphrase'. Below the header, there are two tabs: 'README' (which is selected) and 'License'. The main content area features a section titled 'Introduction in Burmese (Myanmar Language)' with the following text:

ဒီ corpus က paraphrase လိုခေါ်တဲ့ စကားလုံး မတူတာတွေကို သုံးထားပေမဲ့ စာကြောင်း တစ်ကြောင်းလုံးအနေနဲ့ က အမြိုက်အသုံးပြုပြုတဲ့ တူတယ်၊ မတူဘူး ဆိုတာကို ကွန်ပျူးတာက ခွဲခြား သိနိုင်တဲ့ မော်ဒယ်ကို စမ်းဆောက်ကြည့်ဖို့ အတွက် အသုံးပြုဖို့ ရည်ရွယ်ပြီး ဆောက်ခဲ့တဲ့ corpus တစ်ခုပါ။ မြန်မာစာ NLP သုတေသန အလုပ်အတွက် အသုံးဝင်ပါလိမ်မယ်။ ကျွန်တော်နဲ့ ကျွန်တော် Ph.D. ကျောင်းသူ မမြင့်မြင့်ငွေး တို့က ၂၄၀၀ကျော် အချိန်ယူ ပြင်ဆင် ထားခွဲကြတာပါ။ စာကြောင်းရေး စုစုပေါင်း လေးသောင်းလေးရာကျော် ရှိပါတယ်။ open-test data အနေနဲ့လည်း သပ်သပ် စာကြောင်းရေး တစ်ထောင်ကို ပြင်ဆင်ခဲ့ကြပါတယ်။ နောက်ပိုင်း အခြေအနေ ပေးရင်ပေးသလို ဆက်လက်ပြီး corpus ကို တည်ဆောက်သွားဖို့ ရည်ရွယ်ထားပါတယ်။

Versions Information

Version 1.0 Release Date: 3 December 2022

→ <https://github.com/ye-kyaw-thu/myParaphrase>

Corpora (myParaphrase)

Data Format Example

CSV header is as follows:

```
"id","pid1","pid2","paraphrase1","paraphrase2","is_paraphrase"
```

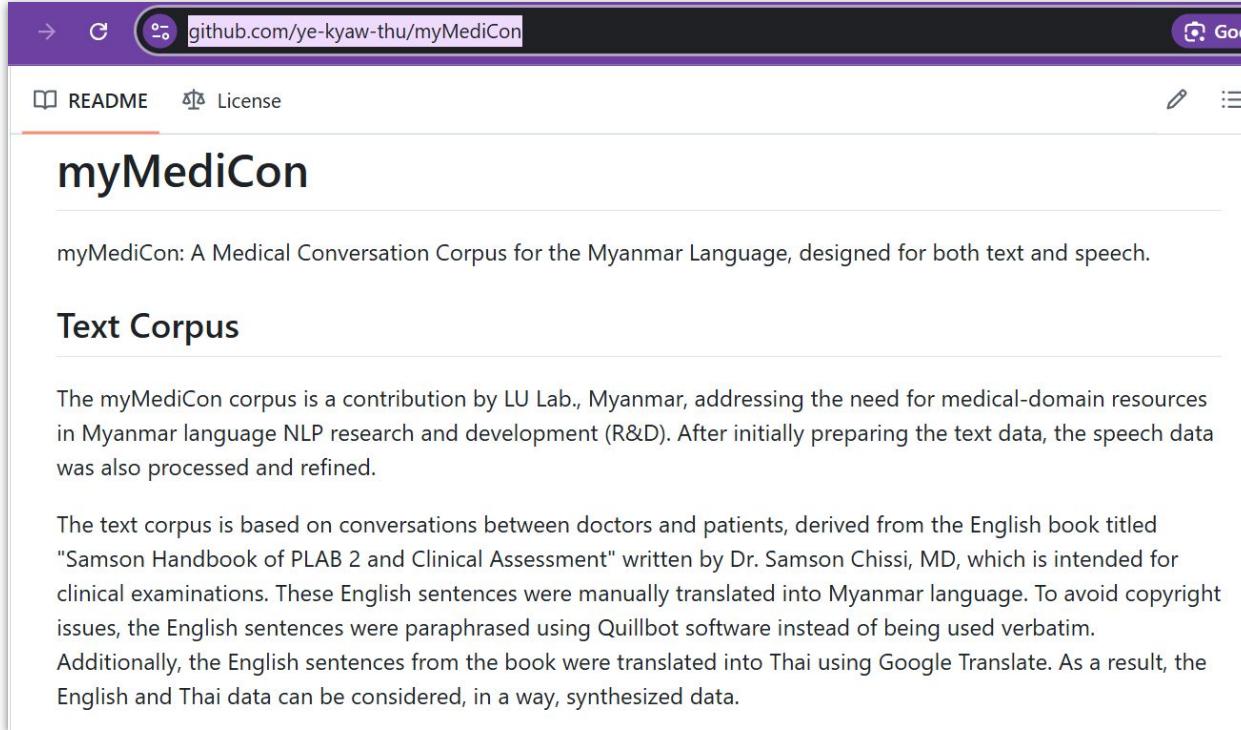
Corpora (myParaphrase)

- "19830","19831","19832","ဒီလို မကြာမကြာ အတူ စား ကြရအောင် ။","ငါတို့ မင်းကို ကျေးဇူးတင် တယ် ။","0"
- "24566","24567","24568","မင်း ဘယ် ရထား စီးလာ မှာလဲ ။","ကျွန်ုပ်မက သက်သက်လွှတ်သမားရှင့် ။","0"
- "28755","28756","28757","ရွှေအေးအေးလေး က အမောပြ စေ တယ် ။","ရွှေအေးအေးလေး က အမော ကု ပြ သွား တာ ပဲ ။","1"
- "23088","23089","23090","ဘယ်သူ ကို သံသယဖြစ် တာ လဲ ။","ကဗျာ ရေးသလား ။","0"
- "16027","16028","16029","တော် ပါ တယ် ကြိုးစား ပါ","တော် လိုက် တာ ကြိုးစား နော် အားမလျှော့ နဲ့","1"

Corpora (myParaphrase)

- myParaphrase ကောပတ်စုနဲ့ ပတ်သက်ပြီး conference စာတမ်းတစ်စွင် နဲ့ ဂျာနယ် စာတမ်း တစ်စွင်စိ ရေးသားခဲ့ကြပါတယ်။
- Myint Myint Htay, Ye Kyaw Thu, Hnin Aye Thant, Thepchai Supnithi, "Statistical Machine Translation for Myanmar Language Paraphrase Generation", In Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2020), Nov 18 to Nov 20, 2020, Bangkok, Thailand, pp. 255-260. [Paper] (**Best Paper Award**)
- Myint Myint Htay, Ye Kyaw Thu, Hnin Aye Thant, Thepchai Supnithi, "Deep Siamese Neural Network Vs Random Forest for Myanmar Language Paraphrase Classification", Journal of Intelligent Informatics and Smart Technology, Oct 2nd Issue, 2022, pp. 25-1 to 25-9. (Submitted Feb 21, 2022; accepted July 17, 2022; published on 31 Oct 2022)

Corpora (myMedicon)

A screenshot of a GitHub repository page for "myMedicon". The repository URL is "github.com/ye-kyaw-thu/myMedicon". The page shows the README and License files. The README file contains the following content:

myMedicon

myMedicon: A Medical Conversation Corpus for the Myanmar Language, designed for both text and speech.

Text Corpus

The myMedicon corpus is a contribution by LU Lab., Myanmar, addressing the need for medical-domain resources in Myanmar language NLP research and development (R&D). After initially preparing the text data, the speech data was also processed and refined.

The text corpus is based on conversations between doctors and patients, derived from the English book titled "Samson Handbook of PLAB 2 and Clinical Assessment" written by Dr. Samson Chissi, MD, which is intended for clinical examinations. These English sentences were manually translated into Myanmar language. To avoid copyright issues, the English sentences were paraphrased using Quillbot software instead of being used verbatim. Additionally, the English sentences from the book were translated into Thai using Google Translate. As a result, the English and Thai data can be considered, in a way, synthesized data.

→ <https://github.com/ye-kyaw-thu/myMedicon>

Corpora (myMedicon)

The screenshot shows the myMedicon application interface. At the top, there is a navigation bar with a file icon, the text "main", and a dropdown arrow. To the right of the dropdown is the text "myMedCon / text / ver0.8 /" followed by a copy icon. Below this is a section titled "ye-kyaw-thu" with the sub-instruction "Add files via upload". A small circular profile picture is next to the name. On the left side, there is a sidebar with a "Name" header and a list of files:

- ..
- data.my
- data.th
- paraphrase.en

- အင်လိပ်စာက မောဒယနဲ့ paraphrase လုပ်ထားတာပါ။
- မြန်မာစာနဲ့ထိုင်းက machine translation နဲ့ ဘာသာပြန်ထားတာပါ
- မြန်မာစာကိုတော့
မဟာတန်းဘဲလုပ်နေတဲ့
ကျောင်းသူ မြှေအစ် (SIIT,
Thailand) နဲ့အတူ
စစ်နိုင်သလောက်
စစ်ထားပါတယ်
- En-Th-My parallel corpus
ပါ

Corpora (myMedicon)

262	မနက်ပိုင်း ဆို ခင်ဗျား ရဲ့ မျက်လုံး တွေ့ မှာ မျက်ဝတ် တွေ့ ရှိ လား ॥
263	မနက်ပိုင်း မှာ ပြည့် ထွေဗ် တာ ပိုခိုး လား ॥
264	ခေါင်းကိုက် တာ နဲ့ ပတ်သက် နေ သလား ॥ ဟုတ်တယ် ဆိုရင် ပထမ ဆုံး အကိုမ် လား ॥ ခေါင်းကိုက် တာ က နေ့ တစ် နေ့ ရဲ့ ပုံမှန် အချိန် တိုင်း မှာ ဖြစ် လား ॥
265	မျက်စိ ထဲ မှာ နာကျင် မူ ရှိ လား ॥ ခင်ဗျား ရဲ့ အမြင် အာရုံ ထိခိုက် လား ॥ (ရေတိမ်)
266	ခင်ဗျား ဖူး နေ တာ လား ॥ ခင်ဗျား အနဲ့ ခဲ့ လား ॥ (ရေတိမ်)
267	ခင်ဗျား မတော်တဆ မျက်လုံး မှာ ထိခိုက် ဒဏ္ဍာ ခံစား ခဲ့ ရ လား ॥
268	ခင်ဗျား အက်စပရင် ဒါမုမဟုတ် ပါဖရင်း ကွဲသို့ သော ဆေး တွေ့ သောက် နေ တာ လား ॥ (မျက်လုံး ရှေ့ပိုင်း အမြှေးပါး အောက် က သွေးကြာ များ သွေးထွေဗ် ခြင်း)
269	အဆစ်မြစ် ကိုက် တာ ရှိ လား ॥ (အဆစ် ရောင် ရောဂါ / သွေးလေးဖက်နာ)
270	ခင်ဗျား ဝမ်းပျော် ဝမ်းလျှော် ဖြစ် ဖူး လား ॥ ခင်ဗျား ရဲ့ မစင် ထဲ မှာ သွေး ပါ တာ သတိစား မိ လား ॥ (အူလမ်း မ ငြိမ် သည့် ဝမ်း ခဏခဏ ပျက် သည့် ရောဂါ)
271	ခင်ဗျား ရဲ့ တစ်ကိုယ်ရည့် အကို က နေ အရည့် ထွေဗ် လား ॥ (ဆီးပြန် ရောင် ခြင်း၊ မျက်စိ နာ ခြင်း နှင့် အဆစ်အမြစ် များ ရောင် ခြင်း တို့ ပါ သော လူငယ် များ ထွေ့ ဖြစ် သည့် လက္ခဏာ စု)
272	ခင်ဗျား ရဲ့ မျက်လုံး ထဲ မှာ သဲတရှပ်ရှပ် ဖြစ် တဲ့ ခံစား မူ ရှိ လား ॥ (ပြင်ပ ဝင်ရောက် လာ တဲ့ အရာ)
273	ကျော် နာ တာ ရှိ လား ॥ ခင်ဗျား ရဲ့ မိသားစု ထဲ မှာ တူဘီး တဲ့ ရောဂါ ရှိ တဲ့ သူ ရှိ လား ॥ (အဆစ်အမြစ် ခိုင် ရောဂါ)
274	အိုင်္ဂီ္ဒီ အစီတေဇေလမိုက် ရှဝဝ် မိလိုကရမ်
275	မျက်စုံး၊ ဘီတာ-ဘလ္ခာကာစ် (တိမိုလော)၊ ဖီလိုကာပိုင်း
276	မျက်စိ အထူးကု ဆရာဝန်
277	ခင်ဗျား က ခေါင်းကိုက် ပြီး ဆေးရုံး ကို ရောက်လာတဲ့ အသက် ၂၀ အရှုယ် အမျိုးသမီး တစ် ယောက် ဖြစ် တဲ့ မစွစ် လိုပတ် ပါ။
278	ပါရာစီတမော သောက် ခဲ့ တယ် ဒါပေမဲ့ အကျိုး မ ရှိ ခဲ့ ဘူး ॥

Corpora (myMedicon)

420	ခင်ဗျား ခေါင်းမှုး နေ လား ဒါမုမဟုတ် ရှိဝေဝေ ဖြစ် နေ လား။
421	ခင်ဗျား ခန္ဓာကိုယ် ထဲ မှာ တစ်နေရာရာ က သွေး ထွက် တာ သတိထား မိ လား။
422	အဆုတ် ပြည့်တည်နာ (ချောင်းဆိုး ပြီး သလိပ် ပါ ခြင်း၊ အဖျား တက် လိုက် ကျ လိုက် ဖြစ် ခြင်း၊ ရင်ဘတ် အောင့် ခြင်း၊ အက်အိုဘီ သွေးချောင်းဆိုး ခြင်း)
423	(ကျူးဗာကုလိုးဆစ်) တိဘီရောဂါ (ကိုယ် အလေးချိန် လျော့ ခြင်း၊ ယူဘက် အေား ပြန် ခြင်း၊ အဖျား၊ များသောအားဖြင့် အာဖရိက ဒါမုမဟုတ် အာရုံ မှ လူနာ ဒါမုမဟုတ် ယူကော မှာ နေ ရင်၊ အရှင်နဲ့ နိုင် တယ်)
424	အဆုတ် ကုန်ဆာ ရောဂါ (အသက်ကြီး လူနာ၊ ကိုယ် အလေးချိန် လျော့ ခြင်း၊ သွေးချောင်းဆိုး ခြင်း၊ မောပန်း ခြင်း၊ အသက်ရှူ။ မ ၁၀ ခြင်း)
425	ဂုတ်ပေါ်ရာ ဆင့်အရှင်း (သွေး ချောင်းဆိုး ခြင်း၊ ဆီး မှာ သွေး ပါ ခြင်း၊ ကဲ့သို့သော ကျောက်ကပ် ပြဿနာ များ ဒါမုမဟုတ် ဆီး တွင် ပရိုတင်းစာတ် ပါ ခြင်း)
426	အေး ပေး ခြင်း (ဝါဘာရင်း)
427	စိတ်ဒဏ်ရာ (ဒဏ်ရာ ဖြစ် ဖူး တဲ့ မှတ်တမ်း)
428	အဆုတ် ထဲ ရှိ သွေးကြော များ တွင် သွေးခဲ့ ပိတ်ဆို ခြင်း (သွေး ချောင်းဆိုး ခြင်း၊ ရင်ဘတ် အောင့် ခြင်း၊ အက်အိုဘီ၊ မိန့်မင်ယ် ဒါမုမဟုတ် မကြာခင် က ခွဲစိတ် မှ လုပ် ထား ခြင်း၊ လေယာဉ် အကြောက်ကြီး စီး ခြင်း၊ ခရာသလုံး နာ ခြင်း၊ ကဲ့သို့သော အန္တရာယ် ဖြစ်နိုင် ခြေ များ တဲ့ အကြောင်းရင်း နဲ့ လူနာ များ)
429	ဘယ်ဘက် သွေးသုံး ခန်း အလုပ် မ လုပ် နိုင် ခြင်း (အိုင်အက်ဒီ ဒါမုမဟုတ် အရှင် က အမ်အိုင် မှတ်တမ်း)
430	လေပြန်၊ အဆုတ်၊ ကျောက်ကပ် တို့ မှ ဖြစ် တဲ့ ခန္ဓာကိုယ် ခုခံ စွမ်းအား စနစ် ယိုယျာင်း မှ ကြောင့် ဖြစ် တဲ့ ရောဂါ (ဆီး တွင် သွေး ပါ ခြင်း၊ သွေး ချောင်းဆိုး ခြင်း)၊ နာရုပ် ယို ခြင်း၊ ကိုယ် အလေးချိန် လျော့ ခြင်း၊ မောပန်း ခြင်း)
431	အဆုတ်ပြုန် များ ဖောင်းပဲ ရောင်ရမ်း ခြင်း (သက်လတ် ပိုင်း လူကြီး၊ ပြည့်တည် နာ၊ ချောင်းဆိုး ခြင်း)
432	အပေါ် လေပြန် လမ်းကြောင်း ရောဂါ (နာရော ခြင်း၊ အအေးပတ် ခြင်း၊ တုတ်ကျေး ကဲ့သို့ လက္ခဏာ များ၊ နာစေး ခြင်း)
433	(နိုးနိုယား) အဆုတ် ရောင် ရောဂါ (အဖျား၊ ချောင်းဆိုး ခြင်း၊ အက်အိုဘီ၊ ရင်ဘတ် အောင့် ခြင်း၊ သလိပ် ထွက် ခြင်း)
434	ဆစ်စာစ် ဖိုက်ပရိုဆစ် (ရင်ဘတ် ပိုးဝင် ခြင်း၊ မှတ်တမ်း နှင့် ကလေး တစ်ယောက် အနေဖြင့် ဖွံ့ဖြိုး မှ နည်း ခြင်း)
435	သွေး ထွက် ပုံမှန် မဟုတ် ခြင်း (သွေးမဲ့ နိုင် သော မျိုးရှိုး လိုက် သည့် ရောဂါ၊ မျှေးရာပါ သွေးယိုစီး မှ ရောဂါ)

→ Medical domain ဖြစ်တာကြောင့် ဘာသာပြန်ရတာ ခက်ခဲပါတယ်

Corpora (myMedicon)

- 420 Do you experience lightheadedness or dizziness?
- 421 Do you have any other areas of your body where you are bleeding?
- 422 Lung abscess (sputum-producing cough, temperature swings, chest pain, SOB, hemoptysis)
- 423 Tuberculosis (weight loss, fever, night sweats, generally from Africa or Asia; if the patient lives in the UK, they are probably an alcoholic)
- 424 Elderly patient with bronchiogenic cancer who exhibits weight loss, hemoptysis, cough, fatigue, and shortness of breath
- 425 Goodpasture's syndrome (hemoptysis, renal issues such as proteinuria or haematuria)
- 426 Warfarin medication
- 427 Trauma (a history of trauma will be present)
- 428 Haemoptysis, chest discomfort, SOB, young female, or any patient with risk factors such as recent surgery, extended travel, or calf pain) are symptoms of pulmonary embolism.
- 429 Left ventricular failure (prior MI or history of IHD)
- 430 Wegener's granulomatosis (excessive sweating, weight loss, rhinorrhea, haematuria, and weariness)
- 431 Middle-aged male with bronchiectasis, persistently purulent sputum, cough
- 432 Upper respiratory tract infection (runny nose, cold, sneezing, flu-like symptoms)
- 433 Pneumonia (fever, cough, sore throat, chest pain, and production of sputum)
- 434 Cystic fibrosis (recurrence of chest infections and underdevelopment during childhood)
- 435 bleeding problems (Von Willebrand disease, hemophilia)
- 436 Equipment (such as a bronchoscopy)
- 437 A history of myocardial infarction, pink, foamy sputum, and shortness of breath when resting flat are signs of pulmonary oedema.

→ අද්ද්‍යාපන උස්ස දුපත

Corpora (myMedicon)

420	คุณ รู รักสี กาก บีบ นศิริช ะหรี หอน นำมี ดหรี อีมี ด?
421	คุณ สังเกตเห็น เลือด ออกจาก ที่อึน ใน ร่างกาย ของ คุ ณหรี อีมี ด?
422	ผ ฝ ไน ปอด (ไอ มี เสมหะ ไข้ แกรง เส็บ หน้าอก SOB ไอเป็น เสือ ด)
423	รักโนร็อก (น้ำหนัก เด หน่องอก ตอนกลางคืน มี ไข้ ผู้ป่วย มาก มาก แผลฟริกา หรือ เอเชีย หรือ ทาง อาศัย อุบัติ สร้างของทางเลือก มีแนวโน้ม ว่าจะ เป็น คน ติดเหล้า)
424	มะเร็ง หลอดลม (ผู้ป่วย สูง อายุ น้ำ ดื่มน้ำ กล ด, ไอเป็น เสือ ด, ไอ, อา บ ล่อน เพ ศ ย, หายใจลำบาก)
425	Goodpastur e's syndrome (ไอเป็น เสือ ด, ปัญหา เกี่ยวกับ ไต เช่น ภาวะ เลือด ซึ่ง หรือ โปรตีน นีน บีส สาวะ)
426	ยา (วาร์ฟาริน)
427	Trauma (จะ มี ประวัติ ของ การ บาด จีบ)
428	เส้นเลือด อุดตัน ที่ ปอด (ไอเป็น เสือ ด, เจ็บ หน้า อก, SO B, หอบ ใจ หายใจลำบาก หรือ ผู้ป่วย ที่มี ปัจจัย เสี่ยง เช่น การผ่าตัด ล่า สุ ด, เที่ยว บิน ยา ว, ปวด น่อง)
429	ภาวะ หัวใจห้องล่าง ขยาย ล้มเหลว (ประวัติ IHD หรือ MI ก่อ หนอง น้ำ)
430	แกรนูลอม่า ตซ ชิส ของ Wegener (เลือด ออก, ไข้ เลือด ออก, น้ำ นูก ไอล, ภาร ณ ตัน ดี หน้า อก, ความ เหนื่อย ล้า)
431	โรค หลอดลม โป่ง พอง (ขยาย รับ กล าง ค น, เสมหะ เป็นหนอง เรื้อรัง, ไอ)
432	การติดเชื้อ ทางเดินหายใจ สาวนะ (จำ บีบ อาการ คล้าย ไข้หวัดใหญ่ น้ำ นูก ไอล)
433	Cystic fibrosis (ประวัติ การติดเชื้อ ที่ หน้าอก ซ้ำ และ ไม่สามารถ เจริญเติบโต ได้ ต่อ มี ไอเป็น เสือ ด)
434	เลือดออก ผิด ปกติ (Haemophili a, Von Willebrand disease)
435	เครื่องมือ วัด (เช่น bronchoscopy)
436	อาการบวม หน้า ที่ ปอด (เสมหะ แพ ง ง รัช มะพุ หายใจ ถ า เมื่อ นอน รา บ, ประวัติ ของ กล้ามเนื้อหัวใจ ตาย)
437	คุณ เคย เดินทาง ไป ต่างประเทศ เมื่อ เร ฯ ๆ ฉ า?
438	คุณ มี อาการ เหื่อ ออก มาก เกิน ไป ใน เก้า กลางคืน หรือ อีมี ด?
439	คุณ มี อาการ ร้อน ร้อน ของ คุ ณหรี อีมี ด?
440	คุณ สังเกตเห็น เลือด ใน กระดอง ร้อน ของ คุ ณหรี อีมี ด?
441	มี ปัญหา อะไร กับ คุณ กระดอง ร้อน ย?
442	มี ปัญหา อะไร กับ คุณ กระดอง ร้อน ย?

→ ထိုင်းခြေတာ ဥပမာ စာကြောင်းတရီး။

Corpora (myMedicon)

- myMedicon ကောပတ်စဲ့ speech data (အသံဒေတာ) ကိုသုံးပြီး လုပ်ခဲ့တဲ့ Automatic Speech Recognition (ASR) ရလဒ်တွေကို အတည်နှင်းမှု ကျင်းပခဲ့တဲ့ LREC-COLING 2024 conference မှာ စာတမ်း ဖတ်ခဲ့ကြပါတယ်။
- Hay Man Htun, Ye Kyaw Thu, Hutchatai Chanlekha, Kotaro Funakoshi and Thepchai Supnithi, "End-to-End Burmese Automatic Speech Recognition for Medical Conversations", the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May, Torino, Italy, pp. 12032–12039

Corpora (MSL4Emergency)

MSL4Emergency

လက်သက်တဲ့ ဘာသာစကားက နားမကြားသူများ၊ အကြားအာရုံချိတဲ့သူများ၊ အပြောဆိုင်ရာမသန့်စွမ်းသူနှင့် နားလည်းမကြား စကားလဲ မပြောနိုင်သူတွေ အတွက် မရှိမဖြစ်လိုအပ်သော ဘာသာစကား တစ်ခုဖြစ်ပါသည်။

သူတိအချင်းချင်း သာမက၊ မိသားစုံ၊ အများနဲ့ ဆက်သွယ်ရာမှာလည်း အလွန်အသုံးဝင်ပြီး၊ ငယ်ရွယ်စဉ်ကလေးဘဝအရွယ်ကတည်းက ပညာရပ်တစ်ခုခုကို လေ့လာ သင်ယူဖို့၊ သင်ကြားပေးဖို့အတွက်လည်း အသုံးပြုနိုင်ပါသည်။ ကျွန်တော်တို့ မြန်မာလက်သက်တဲ့ဘာသာစကားကို လက်ရှိအနေအထားကနေ ပိုမိုဖိုးတိုးတက်ဖို့နဲ့ နားကြားတဲ့သူတွေကြားတဲ့မှာလဲ လေ့လာချင်တဲ့သူတွေက လေ့လာနိုင်အောင် ကွန်ပူးတာနည်းပညာနဲ့ လက်သက်တဲ့ဘာသာစကားကနေ မြန်မာစာကို ဘာသာပြန်ဖို့ အတွက် သုတေသနပြုလျက်ရှိပါသည်။ အဲဒီအတွက် လက်သက်တဲ့ပို့ယိုတွေကို လပေါင်းများစွာ စွောင်းထားတဲ့အထဲက သဘာဝဘေးအန္တရယ်ကျရောက်တဲ့အခါ မှာ အသုံးပြုနိုင်မည့် စာကြားများ၊ အရေးအကြားင်း အခြေအနေမျိုးမှာ နားမကြားသူတွေနဲ့ နားကြားသူတွေရဲ့အကြား ဆက်သွယ်နိုင်ရန်အတွက် လိုအပ်မယ့် အပိုင်းနှုပ်သက်တဲ့ ဖို့ယို ဤရွှေ့င် ကို အားလုံးအတွက် GitHub မှာ ဝေမျှပေးပါသည်။

→ <https://github.com/ye-kyaw-thu/MSL4Emergency>

Corpora (MSL4Emergency)

Some examples of Myanmar written text and Myanmar sign language transcription are as follow:

- မီးချိတ်။ မီးချိတ်
- သဲ အိတ်။ အိတ် သဲ သဲ
- မီးသတ်ဆေးဘူး။ အနီးဘူးဖြန်း
- မီးလောင် လွှယ် သော ပစ္စည်း များ။ မီးလောင် တစ်ခါတည်း ဓာတ်ဆီ စက္က။ စွန့်ပစ် အမျိုးမျိုး
- လောင်စာဆီ။ ဓာတ်ဆီ ၉ ၃ ၉ ၅ အောက်တိန်း အမျိုးမျိုး
- မီးခလုပ်။ ခလုပ် နှိပ်
- မီးခိုး။ မီး အခိုးအငွေ့ အမဲ
- အောက်စီဂျင်။ အောက်ဆီဂျင် ၁၂
- ထွက်ပေါက်။ ဆိုင်းဘုတ် အဲဒီက ပြေး တံခါးပေါက်
- မီးသတ်ကား။ ကား အရေးပေါ်မီးသတ်

Corpora (MSL4Emergency)

The image shows a screenshot of a YouTube playlist page. The URL in the address bar is youtube.com/playlist?list=PLcEIKqU2ZkWKe0aGzGoKH5OtdjSHF3L_f. The left sidebar includes links for Home, Shorts, Subscriptions, and You. The main content area displays a playlist titled "Myanmar-Sign-language-for-..." by Kuma Waseda, which contains 49 videos and has 2,292 views. A "Play all" button is available. The first four videos in the list are shown as thumbnail images of a man in a white shirt performing sign language against a blue background. The details for these four videos are:

- 1 idx20 1 Kuma Waseda • 558 views • 7 years ago 0:03
- 2 idx20 2 Kuma Waseda • 369 views • 7 years ago 0:05
- 3 idx20 3 Kuma Waseda • 145 views • 7 years ago 0:04
- 4 idx20 4 Kuma Waseda • 69 views • 7 years ago 0:04

Corpora (MSL4Emergency)

- မြန်မာ လက်သက္ကတစကား နဲ့ ပတ်သက်တဲ့ စာတမ်းတွေလည်း ရေးသားခဲ့ပါတယ်
- Swe Zin Moe, Ye Kyaw Thu, Hnin Aye Thant, Nandar Win Min, and Thepchai Supnithi, "Unsupervised Neural Machine Translation between Myanmar Sign Language and Myanmar Language", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 53-61. (Submitted December 21, 2019; accepted March 6, 2020; revised March 16, 2020; published online April 30, 2020)
- Ni Htwe Aung, Ye Kyaw Thu, Su Su Maung, Swe Zin Moe, Hlaing Myat Nwe, "Transfer Learning Based Myanmar Sign Language Recognition for Myanmar Consonants", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 35-44. (submitted December 21, 2019; accepted March 6, 2020; revised March 16, 2020; published online April 30, 2020)

Corpora (MSL4Emergency)

- Hlaing Myat Nwe, Ye Kyaw Thu, Hnin Aye Thant, "Myanmar SignWriting Keyboard Mapping Layout for Fingerspelling", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 45-52. (submitted December 21, 2019; accepted March 6, 2020; revised April 25, 2020; published online April 30, 2020)
- Swe Zin Moe, Ye Kyaw Thu, Hlaing Myat Nwe, Hnin Wai Wai Hlaing, Ni Htwe Aung, Khaing Hsu Wai, Hnin Aye Thant, Nandar Win Min, "Development of Natural Language Processing based Communication and Educational Assisted Systems for the People with Hearing Disability in Myanmar", Journal Linguas & Letras, e-ISSN: 1981-4755, ISSN: 1517-7238, Date of Publication online: December/2019, Vol. 20 N. 48 (DOI: 10.5935/1981-4755.20190031)

Corpora (MSL4Emergency)

- Hnin Wai Wai Hlaing, Ye Kyaw Thu, Swe Zin Moe, Hlaing Myat Nwe, Ni Htwe Aung, Nandar Win Min, Hnin Aye Thant, "Statistical Machine Translation between Myanmar Sign Language and Myanmar SignWriting", In Proceedings of the 1st International Symposium on Artificial Intelligence for ASEAN Development (ASEAN-AI 2018), March 26-27, 2018, Phuket, Thailand, pp. 65-72.
- Hlaing Myat Nwe, Ye Kyaw Thu, Hnin Wai Wai Hlaing, Swe Zin Moe, Ni Htwe Aung , Hnin Aye Thant , Nanda Win Min, "Two Fingerspelling Keyboard Layouts for Myanmar SignWriting", In Proceedings of ICCA2018, February 22-23, 2018, Yangon, Myanmar, pp. 290-298.
- တွေ့ကြား စာတမ်းတွေ့လည်း ရှိပါသေးတယ်

Corpora (Muthit Braille)

my: ၁၉၂၀ ပြည့် ကျောင်း သား သ ပိတ် ကြီး တွင် ဦး ဖိုး ကျား
သည် ထင် ရှား သော ခေါင်း ဆောင် တစ် ဦး ဖြစ် သည်။ (“U
Pho-Kyar was a prominent leader in 1920, students'
strike.” in English)

bl:

မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား
မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား
မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား မြန်မာ ဘုရား
မြန်မာ ဘုရား မြန်မာ ဘုရား

Corpora (Muhaung Braille)

my: မြို့ မောက် (မ+ြို့+း+ော+က် မေ+ာ+က်+း)

bl: မြို့မြို့မြို့ မြို့မြို့ (မ+ြို့+ြို့+း မြို့+မြို့)

my: စည်းလုံး (စ+ည်+း+း လုံ+း)

bl: စီးစီး စီးစီး (စ+အီ+း+း စီး+အီ+း)

my: ရှိ သေး (ရ+ြိ+ုံး သေ+း)

bl: ရှိရှိ ရှိရှိ (ရ+အို+း သေ+အို+း)

Corpora (Braille)

- ကျွန်တော်တို့ ဆောက်ထားတဲ့ မှသစ် မျက်မမြင်စာ ကောပတ်စုံနဲ့ ပတ်သက်တဲ့
ထတမ်းက အောက်ပါအတိုင်းပါ
- Zun Hlaing Moe, Thida San, Ei Thandar Phyu, Hlaing Myat Nwe, Hnin
Aye Thant, Naw Naw, Htet Ne Oo, Thepchai Supnithi, Ye Kyaw Thu,
"Myanmar Text (Burmese) and Braille (Mu Thit) Machine Translation
Applying IBM Model 1 and 2", Journal of Intelligent Informatics and
Smart Technology, April 1st Issue, 2021, pp. 18-26. (Submitted February
8, 2021; accepted March 11, 2021; revised April 15, 2021; published
online April 25, 2021)

Corpora (Braille)

- បើពេល CADT ការងារនេះមាន ស្ថិតិថ្មីនៅខាងក្រោម និង វិទ្យាល័យ នៃ សាកលវិទ្យាល័យ នគរបាល កម្ពុជា (Institute of Technology of Cambodia) ការណែនាំនេះមានអត្ថបន្ទាត់ខ្លួន និង ចំណាំ នៅក្នុងពាណិជ្ជកម្ម។
- Kimhuoy Yann, Ponleur Veng, Ye Kyaw Thu, Rottana Ly, "Statistical vs Neural Machine Translations for Khmer Braille", the 13th Conference on Information Technology and its applications (CITA 2024), July 19-20, Danang and Hoi An, Vietnam, pp. 232-243. (Kimhuoy Yann, Ponleur Veng and Ye Kyaw Thu contributed equally to this work as first authors.)
- 19-July-2024: Our paper entitled "Statistical vs Neural Machine Translations for Khmer Braille" achieved "**the 2nd Prize**" at the Workshop on Research and Innovation for Students 2024, Danang, Vietnam.

Corpora (Parallel Corpora)

dw: သယ်ဝယ်သား က လူမြင်: ဟုယ် ॥

bm: ဒီကောင်မလေး က လူလွန်: တယ် ॥

(“The girl is so beautiful” in English)

dw: လတ်ဖတ်ရယ် က ရှိမြင်: ဟုယ် ॥

bm: လက်ဖက်ရည် ချို့လွန်: တယ် ॥

(“The tea is so sweet” in English)

dw: ကောနသား ကောန: မှန်းမှန် သွား ဟုယ် ॥

bm: ကောင်လေး ကောင်: မှန်မှန် တက် တယ် ॥

(“The boy goes to school regularly” in English)

Corpora (Parallel Corpora)

- Some Publications:
- Thazin Myint Oo, Thitipong Tanprasert, Ye Kyaw Thu, Thepchai Supnithi, "Transfer and Triangulation Pivot Translation Approaches for Burmese Dialects," in IEEE Access, vol. 11, pp. 6150-6168, 2023, doi: 10.1109/ACCESS.2023.3236804. (Received 15 October 2022, accepted 27 December 2022, date of publication 13 January 2023, date of current version 20 January 2023.)
- Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, Thepchai Supnithi, "Statistical Machine Translation of Myanmar Dialects", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 14-26. (Submitted December 21, 2019; accepted March 6, 2020; revised March 18, 2020; published online April 30, 2020) [Paper] ("the Workshop Paper Award" at the First Joint Myanmar-Thai NLP/AI R&D Workshop, 2019, Chiang Mai, Thailand)

Corpora (Parallel Corpora)

- Hay Man Htun, Ye Kyaw Thu, Hlaing Myat Nwe, May Thu Win, Naw Naw, "Statistical Machine Translation System Combinations on Phrase-based, Hierarchical Phrase-based and Operation Sequence Model for Burmese and Pa'O Language Pair", Journal of Intelligent Informatics and Smart Technology, Oct 2nd Issue, 2021, pp. 1-9.
(Submitted July 16, 2021; accepted October 13, 2021; revised October 20, 2021; published online October 31, 2021)
- Nang Aeindray Kyaw, Ye Kyaw Thu, Hlaing Myat Nwe, Phyu Phyu Tar, Nandar Win Min, Thepchai Supnithi, "A Study of Three Statistical Machine Translation Methods for Myanmar (Burmese) and Shan (Tai Long) Language Pair", In Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2020), Nov 18 to Nov 20, 2020, Bangkok, Thailand, pp. 218-223.

Embeddings (Visual Display of Word Frequency)

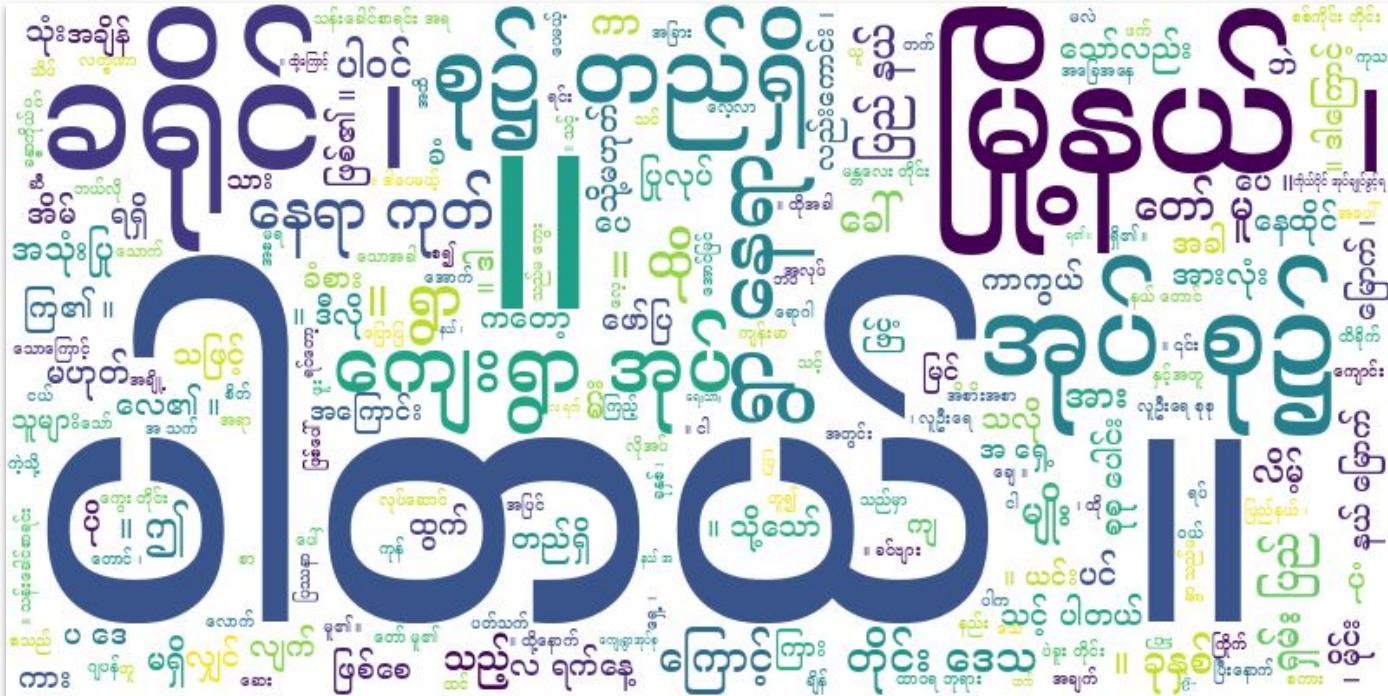


Fig. WordCloud of myMono Corpus version 2.0

Embeddings



Fig. WordCloud of myMono Corpus version 2.0 (removed 100 stopwords + top 1000 words)

Embeddings

$$\text{cosine similarity} = S_C(\mathbf{A}, \mathbf{B}) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}},$$

where A_i and B_i are the i th components of vectors \mathbf{A} and \mathbf{B} , respectively.

Source: https://en.wikipedia.org/wiki/Cosine_similarity

Embeddings (Nearest Neighbor Queries)

Query word? ရခိုင်မှန်တီ

မှန်တီ 0.844894

သက်နှစ်ထမင်း 0.722245

မန်ဟင်းခါး 0.721326

မန်ဟင်းခါးဟင်းရည် 0.700476

မှန်လလုံးရေပြီ 0.68986

ရခုင် 0.687642

အသပ် 0.680808

ပေါင်မန်ကင် 0.672539

အုန်းဆုံးခေါက်ဆွဲ 0.661636

လမှန် 0.659123

Query word? ရခိုင်မှန်တီ

မန်တီ 0.790716

ပေါင်မန်ကင် 0.666917

ရခိုင် 0.638991

ပေါင်မန် 0.635233

မန်ဟင်းခါးဟင်းရည် 0.616322

မှန်လလုံးရေပြီ 0.613742

ကတဗ္ဗန် 0.60678

မန်ဟင်းခါး 0.603243

လမှန် 0.586718

အသားညှပ်ပေါင်မှန် 0.585229

Fig. Nearest neighbor queries with word vectors (left: 100, right: 300 dimensions)

Embeddings (Nearest Neighbor Queries)

Query word? ဆရာမ

ကျောင်းအုပ်ဆရာမကြီး: 0.758842
ကျောင်းသူ 0.720078
ကျောင်းအုပ်ဆရာကြီး: 0.699432
ပညာပြ 0.69658
ဆရာ 0.673178
ကျောင်းအုပ် 0.670142
ကျောင်းသားကတ် 0.669361
ညကျောင်း 0.662287
သုနာပြ 0.661204
ကျိုရင် 0.661188

Query word? ဆရာမ

ကျောင်းအုပ်ဆရာမကြီး: 0.558769
ဆရာ 0.558493
ဆရာလေး 0.556287
ဆရာ 0.505054
ဆရာလှပ် 0.504805
ကျောင်းအုပ်ဆရာကြီး: 0.500963
ဆရာကတော် 0.497177
သုနာပြ 0.495417
ကျောင်းသူ 0.491647
ထသင်တာ 0.489204

Fig. Nearest neighbor queries with word vectors (left: 100, right: 300 dimensions)

Embeddings (Nearest Neighbor Queries)

Query word? ဆရာဝန်

ဆရာဝန် 0.928535

သားဆရာဝန် 0.857536

ခွဲစတ်ကုဆရာဝန် 0.83778

အထူးက 0.790519

ဆရာဝန်ကြီး 0.756213

ဝန်(0.734541

ဝန်/ 0.70739

Laryngoscopy 0.696624

ကုသမ 0.688599

သုနာပြ 0.686658

Query word? ဆရာဝန်

သားဆရာဝန် 0.717381

ခွဲစတ်ကုဆရာဝန် 0.699172

အထူးက 0.65095

ဆရာဝန်ကြီး 0.646915

OG 0.64477

ကုသမ 0.625344

တတေပင 0.613686

ဆေးအညွှန်း 0.602823

cystoscopy 0.594916

Laryngoscopy 0.594098

Fig. Nearest neighbor queries with word vectors (left: 100, right: 300 dimensions)

Embeddings (Word Analogies with FastText)

Query triplet (A - B + C)? ရန်ကုန် မြန်မာ ထိုင်း

ဘန်ကောက် 0.58378

ထိုင်းနိုင်း 0.541044

မြောက်ဥက္ကလာပ 0.535116

ထင်ပေ 0.530853

ချင်းမှုင် 0.508272

မဂ်လာဒံ 0.494238

တောင်ဥက္ကလာ 0.49088

လိုင်သာယာ 0.479702

ထိုင်းဘ 0.467541

သယနှုန်း 0.466632

Fig. (ရန်ကုန် - မြန်မာ + ထိုင်း)?

Embeddings (Word Analogies with FastText)

Query triplet (A - B + C)? စား စားခဲ့ အိပ်ခဲ့

အိပ်ချု 0.663952

အိပ် 0.659455

အိပ်- 0.654276

အိပ်စက် 0.631872

အိပ်ချိန် 0.61765

အိပ်ချင် 0.605428

အိပ်ယာထ 0.564114

အိပ်ပျက် 0.555627

အိပ်၏ 0.551869

အိပ်ယာဝင် 0.541464

Fig. (စား - စားခဲ့ + အိပ်ခဲ့)?

Embeddings (Word Analogies with FastText)

Query triplet (A - B + C)? ကြိုးစား မကြိုးစားဘူး မရှိသေဘူး
ရှိသေ 0.595436
လူရှိသေရှင်ရှိသေ 0.578414
မရှုမသေ 0.566819
မရှုံးမချဖြစ် 0.565208
မထိလေးစား 0.544146
ရှိသေလေးမြတ် 0.543523
မရှုံး 0.538855
ရှိသေကိုင်းညွတ် 0.523898
ကိုးစား 0.49914
နိုင်ထက်စီးနင်း 0.495257

Fig. (ကြိုးစား - မကြိုးစားဘူး + မရှိသေဘူး)?

Embeddings (Word Analogies with FastText)

Query triplet (A - B + C)? ကျောင်းသား ကျောင်း ဘုန်းကြီးကျောင်း
ဘုန်းကြီးကျောင်းသား 0.845826
ဆေးကျောင်းသား 0.717175
နွှဲကျောင်းသား 0.712395
သုံးကျောင်းသား 0.666482
ကျောင်းသားကုဒ် 0.607626
အပောင်းသား 0.566575
နွားကျောင်းသား 0.566006
ကျောင်းသားရေးရာ 0.553581
ရွာဦးဘုန်းတော်ကြီးကျောင်း 0.54597
အကောင်းသား 0.537656

Fig. (ကျောင်းသား - ကျောင်း + ဘုန်းကြီးကျောင်း)?

Embeddings (for Pictures)



pexels-goochie-poochie-3361723.jpg



pexels-lina-1741205.jpg



pexels-pixabay-76957.jpg

Fig. Embeddings for pictures

Language Model

1	6539 → ရွှေ သည် ပဲခူး တိုင်း ဒေသ ကြီး အ
2	4744 → သည် ရွှေ နေရာ ကုတ် မှာ ၂ ၀
3	4672 → သည် ရွှေ နေရာ ကုတ် မှာ ၁ ၈
4	4424 → ရွှေ သည် ရမ်း ပြုပ် နယ် အ ရှေ့
5	4305 → သည် ရွှေ နေရာ ကုတ် မှာ ၂ ၁
6	4087 → သည် ရွှေ နေရာ ကုတ် မှာ ၁ ၇
7	3660 → သည် ပဲခူး တိုင်း ဒေသ ကြီး အ နောက်
8	3511 → သည် ၂ ၀ ၁ ၄ သန်းခေါင်စာရင်း အရ
9	3136 → ဖြစ် သည် ရွှေ နေရာ ကုတ် မှာ ၁
10	3086 → တည်ရှိ သည် ဘူတာ တစ် ခု ဖြစ် သည်
11	3083 → တွင် တည်ရှိ သည် ဘူတာ တစ် ခု ဖြစ်
12	3033 → သည် ပဲခူး တိုင်း ဒေသ ကြီး အ ရှေ့
13	2608 → သည် ရွှေ နေရာ ကုတ် မှာ ၁ ၉
14	2572 → သည် ရွှေ နေရာ ကုတ် မှာ ၁ ၆
15	2529 → ပြုပ် နယ် အ ရှေ့ ကျို့င်း တဲ့ ခရီး
16	2529 → ရမ်း ပြုပ် နယ် အ ရှေ့ ကျို့င်း တဲ့
17	2529 → သည် ရမ်း ပြုပ် နယ် အ ရှေ့ ကျို့င်း
18	2477 → မြို့ တွင် တည်ရှိ သည် ဘူတာ တစ် ခု
19	2191 → ရမ်း ပြုပ် နယ် တောင် လွှိုင် လင် ခရီး
20	2191 → သည် ရမ်း ပြုပ် နယ် တောင် လွှိုင် လင်
21	2185 → ခဲ့ သည် ရွှေ နေရာ ကုတ် မှာ ၁
22	2176 → ရွှေ သည် ရမ်း ပြုပ် နယ် တောင် လွှိုင်

37582	9 → သထံ ခရီး သထံ မြို့နယ် သိမ်ဆိပ် ကျေးရွာ အပ်
37583	9 → သထံ မြို့နယ် ရောင် ဘို့ ကျေးရွာ အပ် စုံ၌
37584	9 → သထံ မြို့နယ် သိမ်ဆိပ် ကျေးရွာ အပ် စုံ၌ တည်ရှိ
37585	9 → သဆ္မာ စ ဒ မဟာ ဓမ္မရာဇာ စီ ရာဇ်
37586	9 → သနုတ်း ရွှေ နေရာ ကုတ် မှာ ၂ ၁
37587	9 → သန္တာသား အနေအထား မုန် အောင် အပိုင် ဖက် မှ
37588	9 → သန်းခေါင်စာရင်း အရ ကျောင်း ကုန်း ကျေးရွာအပ်စုံ တွင် ကျား
37589	9 → သန်းခေါင်စာရင်း အရ သရက် ကုန်း ကျေးရွာအပ်စုံ တွင် ကျား
37590	9 → သပြု သာ ရွှေ သည် မန္တလေး တိုင်း ဒေသ
37591	9 → သဖုန်း ပင် ဆိပ် ကျေးရွာ အပ် စုံ၌ တည်ရှိ
37592	9 → သဘာဝ ဓာတ်ဝန်င်းကြောင်း များ အားလုံး ပါ ရ
37593	9 → သဘာဝ ပစ္စည်း တွေ ကို အကြော် ပြီး ထုတ်
37594	9 → သဘာဝ ပတ်ဝန်းကြောင်း ထိန်းသိမ်း ရေး ဝန်ကြီးဌာန သစ် တော်ဦး
37595	9 → သဘာဝ ပါ ပဲ ရွှေ နေရာ ကုတ် မှာ
37596	9 → သမထ ဘာဝနှင့် စိတ် ပိုပိသနာ ဘာဝနာ စိတ် ထို့
37597	9 → သမ္မတ မြို့မာနိုင်ငံ တော် ပြုပ်ထောင်စု အစိုးရအဖွဲ့ အစဉ်းအဝေး အမှတ်
37598	9 → သမ္မတ ပါ စာ သမ္မတ ကမ္မန် သမ္မတ အာ
37599	9 → သမ္မတ သ မာ စီ သမ္မတ ဘု ဇ
37600	9 → သရက် ရောင်း မြို့နယ် အစ် အစ် ကရင် ပ
37601	9 → သရက် တေားရွာ သည်မ ကျေး တိုင်း ဒေသ ကြီး
37602	9 → သရက် ပင် ကျေးရွာ အပ် စုံ၌ တည်ရှိ သည်
37603	9 → သရ လ အ မြို့ဗျား ဘုရား သ ခင့်

Fig. Ngram counting

Language Model

- N-gram Language Model ဟာ စာကြောင်းတစ်ခုရဲ့ ဖြစ်နိုင်ခြေ (probability) ကို တွက်ချက်ဖို့ အသုံးပြုတဲ့ နည်းလမ်းတစ်ခု ဖြစ်ပါတယ်။ ဒါ Model က စာကြောင်းထဲက စကားလုံးတစ်ခုကို သူရဲ့ အရှေ့က စကားလုံး(များ) ပေါ်မှုတည်ပြီး ခန့်မှန်းလေ့ရှိပါတယ်။
- ဥပမာအားဖြင့် "ဗမာ စကား ပြောတတ် သလား" ဆိုတဲ့ စာကြောင်းတစ်ခုလုံးရဲ့ ဖြစ်နိုင်ခြေကို တွက်ချက်ဖို့အတွက် စကားလုံးတစ်ခုချင်းစီရဲ့ ဖြစ်နိုင်ခြေကို အောက်ပါ formula နဲ့ တွက်လုံ့ရပါတယ်။

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^n P(w_i | w_1, \dots, w_{i-1})$$

Language Model

$$P(\text{မမာ } \theta_{\text{ကာ}} : \text{ပြောတတ် } \psi_{\text{လာ}} :)$$

$$= P(\text{မမာ} | <\text{s}>) \cdot P(\theta_{\text{ကာ}} : \text{မမာ}) \cdot P(\text{ပြောတတ်} | \theta_{\text{ကာ}} :) \cdot P(\psi_{\text{လာ}} : | \text{ပြောတတ်})$$

$$\cdot P(</\text{s}> | \psi_{\text{လာ}} :)$$

Language Model

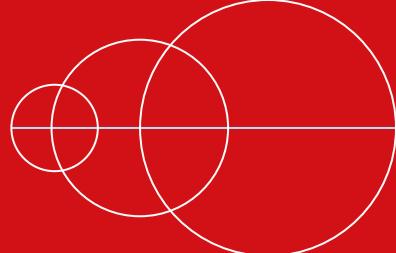
1	\data\	3927799	\3-grams:	11017987	-0.85703826 ଲାଙ୍କି ପାତା ଭୁବେନ୍ଦୁ
2	ngram·1=207187	3927800	-2.0410547 → - - </s>	11017988	-0.6235406 → ଗ ସିଂହ ଭୁବେନ୍ଦୁ
3	ngram·2=3720602	3927801	-0.9211863 → I - </s>	11017989	-0.8632141 → ଅଜ୍ଞା ମାତ୍ରା ଭୁବେନ୍ଦୁ
4	ngram·3=12365250	3927802	-1.9860684 → ଟେ - - </s>	11017990	-0.6254845 → ଶ୍ରୀ ଶ୍ରୀ ଭୁବେନ୍ଦୁ
5		3927803	-1.740716 → ଟ୍ରେ - - </s>	11017991	-0.8766618 → ତ ତଳ୍ପ ଭୁବେନ୍ଦୁ
6	\1-grams:	3927804	-0.12082286 → II - - </s>	11017992	-0.62770605 → ଟ୍ରେ ଶ୍ରୀଶ୍ରୀ ଭୁବେନ୍ଦୁ
7	-6.608784 → <unk> → 0	3927805	-2.0240984 → ମୁ - - </s>	11017993	-1.4885107 → - ଫିର୍ଦ୍ଦୁଗ୍ରୀଯ ଭୁବେନ୍ଦୁ
8	0 → <s> → -1.4534098	3927806	-1.9391153 → ଫ୍ରେଣ୍ଟ - - </s>	11017994	-1.3004675 → ଏର୍ତ୍ତର ଆଗ୍ନିରୂପାତ୍ମିକ ଭୁବେନ୍ଦୁ
9	-2.9836152 → </s> → 0	3927807	-2.4646177 → ଲୟାନ୍ ଭୁବେନ୍ଦୁ	11017995	-1.4840224 → ସି ଶ୍ରୀ ଭୁବେନ୍ଦୁ
10	-2.9946578 → - → -0.48582184	3927808	-1.1716466 → ଲ୍ୟାନ୍ ଭୁବେନ୍ଦୁ	11017996	-0.82077515 → ଏନ୍ଦିଗା ଭୁବେନ୍ଦୁ
11	-3.8384993 → ଲେଖକଙ୍କଣିକା → -0.8468338	3927809	-1.8926232 → ପି - - </s>	11017997	-0.82077515 → ଗ ଏନ୍ଦିଗା ଭୁବେନ୍ଦୁ
12	-2.9895 → ଧର୍ମ → -1.0622561	3927810	-2.0728502 → ଟେଟ୍ରୋ - - </s>	11017998	-0.5713608 → ଅତି ଏନ୍ଦିଗା ଭୁବେନ୍ଦୁ
13	-2.7952466 → ଶ୍ରୀଶ୍ରୀ → -1.1160072	3927811	-1.6670866 → ଅପି - - </s>	11017999	-0.6238863 → ଫିର୍ଦ୍ଦୁଗ୍ରୀଯ ଅଥିଚେ ଭୁବେନ୍ଦୁ
14	-2.1484232 → ଆମ୍ବା → -0.91495144	3927812	-2.3193204 → ଆମ୍ବା - - </s>	11018000	-0.61718965 → ମୁ Animation ଭୁବେନ୍ଦୁ
15	-4.6132345 → ଆମ୍ବାମା → -0.449102	3927813	-2.5437999 → ଆମ୍ବା - - </s>	11018001	-0.34802505 → ବୈରିଟା ମୁତ୍ତିଶ୍ରୀମି ଭୁବେନ୍ଦୁ
16	-2.826191 → ହାତ → -0.64772815	3927814	-1.1367344 → ଗ - - </s>	11018002	-0.35268822 → ଫିର୍ଦ୍ଦେ ତଠି” ଭୁବେନ୍ଦୁ
17	-3.816531 → ଡେବାର୍କିନ୍ଦ୍ରିୟ → -0.6306369	3927815	-1.508018 → ଟେଟ୍ରୋ - - </s>	11018003	-0.7618375 → ଫିର୍ଦେ ତଠି” ଭୁବେନ୍ଦୁ
18	-4.485493 → ଶିଥା → -0.54055643	3927816	-2.6039376 → ଟେଟ୍ରୋ - - </s>	11018004	-1.1923101 → <S> ରେଣ୍ଡିଶନ୍ ଭୁବେନ୍ଦୁ
19	-2.9791422 → ଟେଟ୍ରୋ - - </s>	3927817	-2.255387 → ଏଣ୍ଟିକା - - </s>	11018005	-0.6090044 → ତବା ଫୁର୍ତ୍ତିପିଲି ଭୁବେନ୍ଦୁ
20	-3.0641553 → ଧିନ୍ ଭୁବେନ୍ଦୁ - - </s>	3927818	-1.9319197 → ଧିନ୍ ଭୁବେନ୍ଦୁ - - </s>	11018006	-1.7485399 → ଏଇ ହାତ ଭୁବେନ୍ଦୁ
21	-2.5479517 → ଶ୍ରୀଶ୍ରୀମାତ୍ରା - - </s>	3927819	-1.8834062 → କାମ୍ବିନ୍ ଭୁବେନ୍ଦୁ - - </s>	11018007	-0.6221347 → sympathetic system ଭୁବେନ୍ଦୁ
22	-4.1072397 → ଗୁର୍ବିତର୍କା - - </s>	3927820	-1.8680185 → ଲ୍ୟାନ୍ ଭୁବେନ୍ଦୁ - - </s>	11018008	-0.6221347 → Parasympathetic system ଭୁବେନ୍ଦୁ
23	-4.461196 → କୌଣସିକା - - </s>	3927821	-2.0713744 → ଫେର୍ନ ଭୁବେନ୍ଦୁ - - </s>	11018009	-1.0945097 → <S> ଲାଗନ୍ତି ଭୁବେନ୍ଦୁ

Fig. Ngram language model with ARPA format

Practical NLP Applications

- Monolingual, parallel, tagged corpus တွေနဲ့ ဆောက်ထားတဲ့ Pretrained Model တွေကိုသုံးပြီး
 - ◆ Text Generation
 - ◆ Summarization
 - ◆ Translation
 - ◆ NER Tagging
 - ◆ Sentence Similarity အလုပ်တွေကို လုပ်ပြပါမယ်
- ပြီးတော့ SymSpell Approach နဲ့ မြန်မာစာလုံးပေါင်းအမှားတွေကို ရှာဖွေတာ၊ အမှာန်ပြင်တာတွေကို ဘယ်လို လုပ်လို ရသလဲ ဆုံးတာကိုလည်း သင်ပေးပါမယ်
- လက်ရှိ NLP task တွေကို ဘယ်လောက်ထိ လုပ်ပေးနိုင်သလဲ၊ ဘယ်လို အခက်အခဲတွေ ရှိနေသေးလဲ ဆုံးတာတွေကို လွှဲလာကြရအောင်

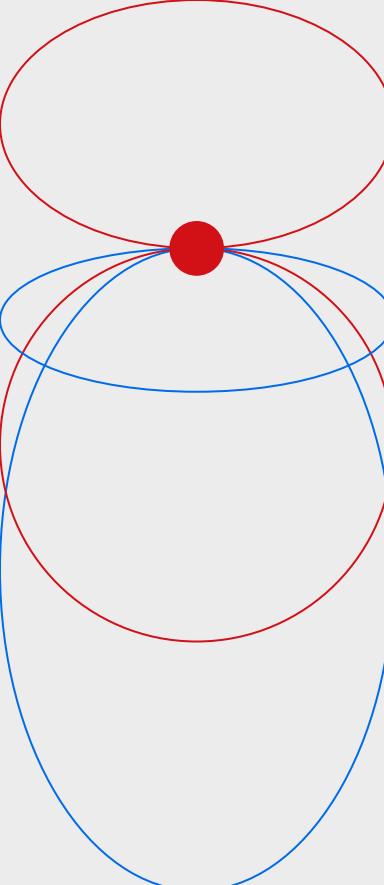
Open Discussion



မေးခွန်းတွေ ရှိရင် မေးကြပါ

Conclusion & Open Discussion

- Pretrained model တေနဲ့ ကျို့စွဲတော်တို့ မြန်မာစာနဲ့ ပတ်သက်တဲ့ အလုပ်တွေ အများကြီးလုပ်လို့ ရနိုင်ပါတယ်
- လက်တွေ့ လုပ်ပြုစဉ် အခါမှာလည်း မြင်ကြရတဲ့ အတိုင်းပါပဲ၊ မြန်မာစာအတွက် မြန်မာတွေ ကုယ်တူင် လုပ်ဖို့ လုအပ်ပါတယ်
- လုအပ်ချက်အရေရာ၊ ဂုဏ်သံက္ခာအရေရာပေါ့
- အထူးသဖြင့် တိုင်းရင်းသား ဘာသာစကားတွေနဲ့ ပတ်သက်ရင် ဒေတာပြင်တဲ့ အပိုင်းအပါအဝင် လုပ်စရာတွေ အများကြီးကျို့နေပါသေးတယ်
- ပြီးတော့ NLP သမားတွေနဲ့ပဲ မဟုတ်ပဲ မြန်မာစာပညာရှင်တွေ၊ ဘာသာမေဇ ပညာရှင်တွေနဲ့ အတူတူ လုပ်ကြရမယ်လို့ ကျို့စွဲတောကတော့ နားလည်ပါတယ်
- စိတ်ဝင်စားတဲ့ ကျောင်းသား၊ ကျောင်းသူတွေရှုံးရင် LU Lab ကို internship အနေနဲ့ join တာပဲ ဖြစ်ဖြစ်လုပ်ကြပါ
- စာလည်း ပိုသင်နိုင်အောင် အချိန်လုပါမယ်



*Thank
you*