Dependency Annotated Dataset of The Myanmar Language

Nwe Nwe Win
Department of Myanmar Nationalities' Languages
Ministry of Education
Naypyitaw, Myanmar
nwenwewin1@ucsy.edu.mm

Win Pa Pa
Natural Language Processing Lab
University of Computer Studies, Yangon,
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract- **Data Annotation is the process of labeling individual elements of training data to help machine translation. This paper presents a new Part of Speech tagset of the Myanmar language using TagEditor tool software to get an annotated dataset of the Myanmar language. The annotated dataset of the Myanmar language can be used to train machine learning models as a TagEditor tool, UD Pipe. My POS is similar to U-POS for the Myanmar language to get the MLPOS Tagset. Dependency structure is a linguistic principle that grammatical processes function primarily on structures in sentences. It presents the preliminary result of the annotated tree structure. This paper intends to be achieved better-quality Annotated Dataset with the use of MLPOS. Larger Dataset will be utilized to get better quality of Data Accuracy. Utilized Sentences include not only simple sentences but also complex sentences. Morphology-enriched Myanmar sentences will be developed Parsing for Morpheme Level. In this paper, over 25,000 sentences will be annotated to develop 700,000 POS Tagging. Parsing experiments are evaluated by UDPipe in terms of unlabeled and labeled attachment scores: (UAS) and (LAS), which are 91.42%, and 86.34% in test experiment respectively.**

Keywords: Dependency Structure, Annotated Dataset, Dependency Tree, Morphology-enriched

I.NTRODUCTION

Dependency parser tools can be utilized to analyze sentences, with a particular focus on their grammatical structure; they pass each sentence into a set of rules which then allow us to determine the dependencies structure and analyze them. After that, the Dataset will be created with the POS tagset. Natural Language Processing (NLP) is a field of computer science that deals with the interaction between computers and humans using natural language, and Parsing is one of the fundamental tasks in NLP. It is the process of analyzing a sentence into its constituent parts and describing their grammatical roles.

The Research work [1] has been done on Myanmar morphological analysis tokenization and part-of-speech (POS)Tagging. It used twenty thousand Myanmar sentences describing in details with the preparation, refinement, and features of the annotated Corpus. A sequence labeling approach of Conditional Random Field was applied in this Tagset were applied to determine part-of-speech. It designed a unified NOVA (9) framework for annotating low resources but high analytic languages.

The Research work [2] is a dependency head annotation scheme with Universal part-of-speech and Universal Dependencies for Myanmar Dependency Treebank. It used 22,810 sentences and 680,218 tokens for annotating from three corpora for Myanmar Dependency Treebank. Some language-specific issues are also described with examples. Annotated Treebank Data have been evaluated by the CoNLL-U 2017 evaluation script for parsing performance. This proposed method is a suitable way for building Myanmar Dependency Tree. Moreover, this dependency head annotation for dependency Treebank is the first work for the Myanmar language.

The research work [3], mentioned in the manual, provides the detailed guidelines for the surface annotation of the Myanmar texts in Asian Language Treebank (ALT). In this paper, the tokenization and part-of-speech (POS) annotation was handled uniformly under an annotation system, named NOVA.

The Research work [4] presented an annotation procedure by unsupervised approach and parsing experiments on the annotated corpus. This research paper describes dependency structures based on language and tagset properties of dependency trees and approaches using raw corpus. It is used U-POS, UDPipe to annotated data. This paper expresses the Mapping Universal POS and Language POS. It also presents the Dependency structure of Noun, Proper Noun, Possessive Noun, Nominal Noun, Adjectives, Conjunctions, Coordinate, Subordinate, and Adverb. Its Parsing experiments are evaluated by UDPipe, (UAS) and (LAS) attachment scores are 93.20% and 91.21%.

MYANMAR LANGUAGE

The Myanmar language, also known as Burmese, is overwhelmingly monosyllabic where one or more monosyllabic morphemes can be combined with different levels of strength. Words combined with multiple morphemes include a stem with zero or more affixes to form a meaningful unit. Affixes, which are mainly suffixes play a prominent role in the grammar of Myanmar because they carry almost all the grammatical information contained in a sentence.

For the annotation of sentences from above-mentioned articles, words and / or phrases must be divided in accordance with the Part of Speech (POS) of the Myanmar Language. In this paper, for the construction of Dependency Corpus, MLPOS will be divided as follows:-

TABLE-1 UNIVERSAL POS AND LANGUAGE POS

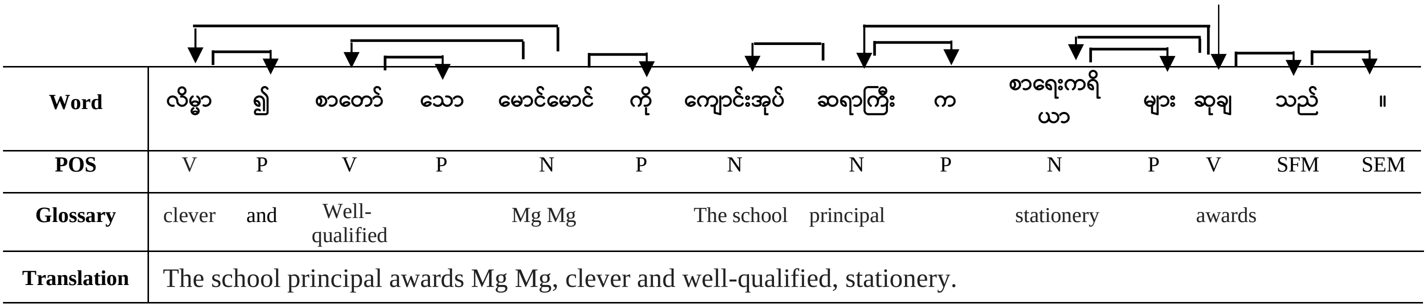
U-POS	General POS	POS of MLPOS	Description / Translation
NOUN	N	N	Noun (တက္ကသိုလ်) (University)
		Nb	Text Number / Number တစ် (one)
		N	Foreign Word (ကား) (Car)
PROPN	N	N	Proper Noun (နေပြည်တော်) (NayPyiTaw)
	ABB	N	Abbreviation (ဂျီဒီပီ) (GDP)
PRON	PRON	N	Pronoun (သူ၊ သူမ) (He, She)
NUM	NUM	Nb	Number [တစ် [one]/ ၁ (1)]
ADJ	ADJ	Adj	Adjective (ဖြူ) (White)
ADV	ADV	Adv	Adverb (လှိုက်လှိုက်လှဲလှဲ) (heartily)
VERB	V	V	Verb (စားသည်) (eat)
CCONJ	CONJ	P	Coordinate Conjunction (နှင့်) (and)
SCONJ	CONJ	P	Subordinate Conjunction (သောအခါ) (When)
PART	PART	P	Particles (များ) [Plural] marker, [Specified Marker]
		P	Negative Particle (မ [not])
ADP	PPM	PPM	Post Position Marker (သည်၊ ကို၊ အား) Norminal Marker (၌၊ မှာ) (at / in) calle as “Adoposition” in universal POS Tag
PUNCJ	PUNC	Pun	Punctuation (။ [.]) ၊ (၊ [,])
INTJ	INT	Int	Interjection (အိုး [oh])
SYM	SB	Sb	Symbol (%, =)

STEP 1. DATA COLLECTION AND DATA PREPARATION

In this first step, we have to consider the data to construct Dependency Corpus and to parse the sentences. The books such as “Dictionary of Myanmar Proper Name (Arts & Literature)” published by the Department of Myanmar Nationalities’ Languages, “The Biography of Myanmar Writers” by Sayargyi Professor U Pe Maung Tin, “Chindwin-Myit-Thar Mu-Ayar” by Sayargyi Dr. Toe Hla, and “Collections of Ethnic and Aesthetics Articles, Stories to educate for well-behavior” were used as the data for parsing the sentences, POS Tagging, UD Tree , CoNLL-U Format and Dependency structure to get Dependency Annotated dataset of the Myanmar Language. Dependency parsing involves exploring the dependencies between words in a sentence to gain an understanding of its grammatical structure. It breaks sentences into multiple components and works on the concept that there are direct links (or dependencies) between every linguistic unit in a sentence. Relations between linguistic units or words are indicated with directed arcs in a typed dependency structure. Relationships between words are indicated by dependency tags. When there are dependencies between two words, one word is the head while the other one is the dependent (or child). Dependency parsing can identify the subjects and objects of a verb, while also showing you which words modify or describe the subject.

STEP 2. PARSING OF SENTENCE

In this paper, to develop Myanmar Dependency Parsing in accordance with POS level, the sentences are chosen from the selected stories. First of all, the selected sentences will be placed in the TagEditor document. For MLPOS Tagging, the part of speech will be defined such as N, V, Adj, Adv, Nb, P, PPM, SFM, etc. By utilizing TagEditor Tool Software, the words in each sentence will be tagged with the use of MLPOS. For the POS Tagging, it needs to check whether there are extra spaces among words in a sentence to be a meaningful sentence and to name one serial number for one sentence. If the sentences are dispersed, the new sentences are adjusted by Uncheck Option. If there are more spaces among sentences, Delete Token option is used for these spaces. To edit the Parsing part, Edit Token is used to correct the mistake words; Insert Token is used for the supplementation. After checking the POS Tagset in accordance with the number of sentences and Word Level, POS Tagging is developed. If the correct Myanmar Language Annotated Dataset is achieved, it will be saved. For the POS attached sentences are used for Dependency Head Annotation; that is, defining Dependent and Independent words for Dependency Parsing. A Dependency Corpus will be constructed with the following example complex sentence လိမ္မာ၍ modified the object phrase မောင်မောင်ကို. This clause is embedded in the object phrase also called the noun phrase.



Word	လိမ္မာ	၍	စာတော်	သော	မောင်မောင်	ကို	ကျောင်းအုပ်	ဆရာကြီး	က	စာရေးကရိ ယာ	များ	ဆုချ	သည်	။
POS	V	P	V	P	N	P	N	N	P	N	P	V	SFM	SEM
Glossary	clever	and	Well-qualified		Mg Mg		The school	principal		stationery		awards		
Translation	The school principal awards Mg Mg, clever and well-qualified, stationery.													

FIGURE 1. DEPENDENCY PARSING OF MYANMAR SENTENCE

TABLE-2 STATISTICS OF DATASET OF MYANMAR LANGUAGE

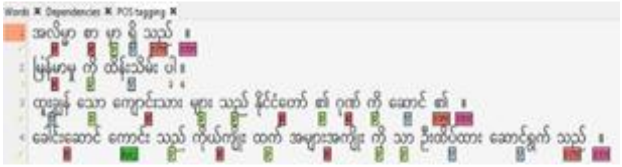
Data Set	Sentences	Word Tokens
Stories to educate for well-behavior	5,373	80,979
Dictionary of Myanmar Proper Name (Arts & Literature)	2,515	35,696
Chindwin-Myit-Thar Mu- Ayar	559	11,516
Collections of Ethnic and Aesthetics Articles	22	363

TABLE-3 ACCURACY OF DEPENDENCY PARSING

MODEL	MLPOS	
	UAS(%)	LAS(%)
UD PIPE	91.42	86.34

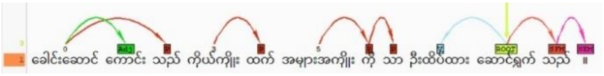
CONCLUSION

This paper has proposed an annotated dataset for Dependency Parsing of Myanmar Language. TagEditor tool software is applied for the MLPOS Tagset, For Dependency Parsing, it will consider carefully that whether which word is dependent on another word. After word, the root dependency relationship will be modified on the root word. This paper presents simple sentences and complex sentences. It includes a Morphological rich sentence. This paper shows present the Dependency of Myanmar Language annotated dataset including over 25,000 sentences and above 700,000 words tokens using Annotated Tools, such as CoNLL-U Format, UDPipe Tools, UD Annotated Dataset, ALT Project, etc. Furthermore, it will be useful tool and Dependency Corpus using related annotated datasets of the Myanmar Language to get high-quality human-powered data annotation. The Universal Dependencies scheme maps all language-specific part-of-speech tags. The Universal POS tags are used in Universal Dependencies, which is a project developing cross-linguistically consistent treebank annotation for many languages. This research paper intends to be useful for people who are interested in NLP and to upgrade the Natural Language Processing into Natural Langue Understanding step by step in the Myanmar language.



1 အလိမ္မာ စာ မှာ ရှိ သည် ။
2 မြန်မာမှု ကို ထိန်းသိမ်း ပါ ။
3 ထူးချွန် သော ကျောင်းသား များ သည် နိုင်ငံတော် ၏ ဂုဏ် ကို ဆောင် ၏ ။
4 ခေါင်းဆောင် ကောင်း သည် ကိုယ်ကျိုး ထက် အများအကျိုး ကို သာ ဦးထိပ်ထား ဆောင်ရွက် သည် ။

FIGURE 2 PART OF SPEECH (POS) TAGSET OF THE SENTENCE



1 ခေါင်းဆောင် ကောင်း သည် ကိုယ်ကျိုး ထက် အများအကျိုး ကို သာ ဦးထိပ်ထား ဆောင်ရွက် သည် ။

FIGURE 3 DEPENDENCY RELATION OF THE SENTENCE

Myanmar Word	မာရဝိဇယ ဗုဒ္ဓ ရုပ်ပွားတော် ၏ ဉာဏ်တော်အမြင့် မှာ ၆၃ ပေ ရှိ ပြီး ပလ္လင်တော်အမြင့် မှာ ၁၈ ပေ ရှိ သည် ။
English glossary	Māravijaya Buddha marble statue of height 63 feet and Throne of height 18 feets
Translation	Māravijaya marble statue of the Buddha is 63 feets and Throne is 18 feets in height

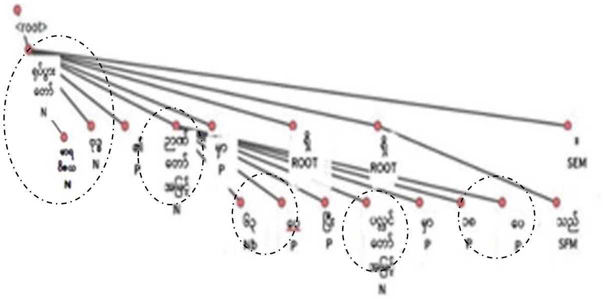


FIGURE 4 UD TREE OF MYANMAR LANGUAGE

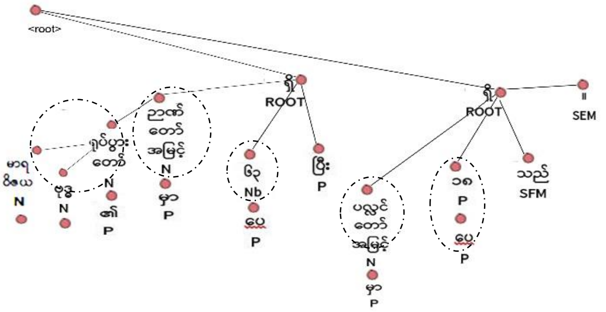


FIGURE 5 REFERENCE DEPENDENCY TREE OF MYANMAR LANGUAGE

REFERENCE PAPER

[1]. CHENCHEN DING, ASTREC, National Institute of Information and Communications Technology, Japan HNIN THU ZAR AYE, WIN PA PA, KHIN THANDAR NWET, and KHIN MAR SOE, "Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-speech Tagging"

[2]. Hnin Thu Zar Aye*, Win Pa Pa "Dependency Head Annotation for Myanmar Dependency Treebank"

[3] Chenchen Ding1, Hnin Thu Zar Aye1,2, Masao Utiyama1, Win Pa Pa2, Eiichiro Sumita1 "Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese)"

[4]. Hnin Thu Zar Aye, Win Pa Pa, Ye Kyaw Thu "UNSUPERVISED DEPENDENCY CORPUS ANNOTATION FOR MYANMAR LANGUAGE"