Proposed Framework for Stochastic Parsing of Myanmar Language

Myintzu Phyo Aung, Ohnmar Aung, Nan Yu Hlaing
Myanmar Institute of Information Technology, Myanmar
myitzu.mm@gmail.com,mamalay2009@gmail.com, nanyuhlaing@gmail.com

Abstract. Parsing is breaking a sentence into its constituent nonterminal. Parsing is useful in the study of artificial intelligence for various reasons, such as, for an index-term generation in an information retrieval; for the extraction of collocation knowledge from large corpora; development of computational tools for language analysis. In this paper, a framework for stochastic parsing of Myanmar Language is presented. This parsing system will use the context free grammar and stochastic context free grammar. Myanmar sentence will be accepted as input. And then this input sentence will be tokenized, segmented and assigned part of speech tags. Finally, stochastic parse tree will be generated by using stochastic context free grammar as the output of stochastic parsing system.

Keywords: parsing, stochastic parsing, stochastic context free grammar, parse tree.

1 Introduction

Text classification or categorization includes automatic classification of documents or texts into predefined categories. Different application of text classification includes spam filtering, email routing, language identification, genre classification, readability assessment etc. Syntactic parsing is a central task in natural language processing because of its importance in mediating between linguistic expression and meaning. One important way in which languages differ from each other is in their syntactic structure. Many languages can be classified in this way; for example, English exhibits SVO order (subject-verb-object), Myanmar Language has SOV (Subject Object Verb) order and so on. Looking beyond the order of the subject, verb, and object, we can also consider the order of other syntactic elements, such as the order of a head noun and its modifiers.

Language is an important role in human communication because it is used not only for expressing thoughts but also for exchanging information. Processing of naturel language is branch of linguistics, artificial intelligence and computer science and its purpose is to have interaction among natural language of human beings and computers. In language processing, examples of such tasks include part-of-speech tagging, named entity recognition. The task which will be focused in this paper is shallow

parsing. Shallow parsing identifies the non-recursive cores of various phrase types in text, possibly as a precursor to full parsing or information extraction.

2 Related Work

In recent years, the artificial intelligence community has studied various stochastic behaviors of natural language (NL) to carry out successful information extraction processes. This section describes the related work of the system. Automatic summarization of Myanmar text was presented in [17]. In their approach, Myanmar document is accepted as input. Firstly, sentence boundary is identified, and word segmentation, stop word removal and part of speech tagging are performed as preprocessing steps. Features such as location, similarity to the title and numerical data in sentence are extracted from sentence by using a training corpus of document summary pair.

Chunking can be used in several tasks and also a previous step providing input to the next steps. A chunk is basically the identification of parts of speech like short phrases and consists of single content word surrounded by a constellation of function words. Myanmar Phrase identification system by using Conditional Random Fields was presented in [20]. The system was firstly trained with training data and tested with testing data. Although there were some error identifications, the system got the best accuracy.

Morphological analysis is an important first step in many natural language processing tasks such as parsing, machine translation, information retrieval, part of speech tagging among others. A probabilistic language model for Joint Myanmar Morphological Segmentation and Part of Speech Tagging was presented in [1]. In their experiments, three testing corpora are used for evaluation in order to calculate the accuracy of the word segmentation and tagging. Each corpus contains 150 sentences these are from news websites and Myanmar Grammar books.

Construction of Finite State Machine (FSM) for Myanmar noun phrase structure in Context Free Grammar is presented by [9] to apply in Myanmar noun phrase identification and translation system which is part of Myanmar to English machine translation system. New phrase chunking algorithm for Myanmar Natural Language processing has been reported in previous study [10].

Phrase Structure Grammar for Myanmar Noun Phrase Extraction is presented in [11]. In their approach, Context Free Grammar is generated and used for extraction of Myanmar Noun Phrase. The system is tested on 1500 testing sentences: 1000 simple sentences and 500 compound sentences. The evaluation measures are defined in terms of the precision, recall and f-measure and shown that the system got the high accuracy.

3 Lexical Categories of Myanmar Language

Myanmar language, also known as Burmese is the official language of the Union of Myanmar. It is spoken by 32 million as a first language and as a second language by ethnic minorities in Myanmar. Burmese is a member of the Tibeto-Burman languages,

which is a subfamily of the Sino-Tibetan family of languages. Burmese is a tonal and analytic language using the Burmese script. This is phonologically based script adapted from Mon, and ultimately based on an Indian prototype. A basic unit of every language is its sentence. Sentences in Myanmar language consist of grammatically structured words. Linguists have grouped these Myanmar words into classes that show similar syntactic behavior and often a typical semantic type. These word classes are known as parts-of-speech (POS). There are eight basic parts-of-speech in the Myanmar language. The two most important POS are noun and verb. These word classes or POS are shown in table 1.

Table 1. Lexical Categories of Myanmar Language

Lexical Category	Definition	Examples
Noun	Words that refer to the names of persons, places, things or concepts	မောင်မောင်၊ ကျောင်းသား၊ စာအုပ်
Verb	Words that express the actions or states in a sentence	စားသောက်
Pronoun	Words that can replace nouns or can be used instead of nouns	ကျွန်တော်၊ကျွန်မ
Adjective	Words that describe the properties of nouns	တော် ၊ထိုး၊ကျ်
Adverb	Words that modify verbs, adjectives and other adverbs	မြန်မြန်၊ခင်ခင်မင်မင်
Preposition	Words that indicate the relationships between the nouns or pronouns and some other parts of speech	သည်၊က၊မှာ
Conjunction	Conjunctions join clauses, parallel nouns, and adjectives	နှင့်၊သို့မဟုတ်၊မှတစ်ပါး
Particles	Words that indicate the subjects, objects, place, or time etc	များ၊တို့

4 Different Levels of Language

In order to understand language, language is studied at different levels. There are seven levels for linguistic analysis. These levels of language study are presented in Table 2. [8]

Table 2. Different Level of Language Study

Level	Description	Example usage/System

Phonological	the interpretation of speech sounds within and across words	speech-recognizing systems
Morphology	the study of the meaningful parts of words	automatic stemming, truncation or masking of words
Lexicology	the study of words	parts-of-speech tagging or the use of lexicons
Syntactic	the study of the rules, or "patterned relations", that govern the way the words in a sentence are arranged	parsing algorithms
Semantics	the study of the meaning of word more complex level of linguistic analysis	identification of automatically phrases of two or more words that when looked at separately have quit e different meanings
Discourse Analysis	works with units of text longer than a sentence	by understanding the structure of a document, Natural Language Processing systems can make certain assumptions
Pragmatics	the study of how the context influences meaning	

The above levels of linguistic processing reflect an increasing size of unit of analysis as well as increasing complexity and difficulty as we move from phonological level to pragmatics.

5 Proposed Stochastic Parsing System

Everyone who has studied a foreign language knows that a grammar is a book of rules and examples which describes and teaches the language. Good grammars make a careful distinction between the sentence / word level, which they often call syntax or and the word/letter level, which they call grammar. A grammar can be used to derive symbol strings having a certain structure. Given an input text and a grammar, whether the text could have been generated by the grammar and how it was generated, that is which grammar rules and which alternatives were used in the derivation. This process is called parsing or syntax analysis. An attempt to reconstruct the derivation, the reconstruction is performed by working from the string symbol down towards the given string. This method is called top down parsing. Bottom up parsing works in the opposite direction from top-down. A bottom-up parse starts with the sting of terminals itself and builds from the leaves upward, working backwards to the start symbol by applying the productions in reverse. This section describes the proposed stochastic parsing system for Myanmar Language which is depicted in Figure 1.

5.1 Preprocessing

Myanmar Language likes other South Asia languages do not place spaces between words but spaces are usually used to separate phrases. Besides, in Myanmar sentences, some are written with propositions or particles and some are written without propositions or particles.

The language is classified into two categories. One is formal, used in literary works, official publication, radio broadcasts, and formal speeches. The other is colloquial, used in daily conversation and spoken. In Myanmar script, sentences are clearly delimited by a sentence boundary marker but words are not always delimited by spaces. Therefore some preprocessing steps such as tokenization, segmentation, and part of speech tagging are required before the parse tree of the sentence is generated.

5.1.1 Tokenization

Tokenization is the process of breaking up the given text into units called tokens. Tokens may be words or number or punctuation mark. Tokenization does this task by locating word boundaries. So tokenization can also be said to be the identification of syllable boundaries. Ending point of a word and beginning of the next word is called word boundaries. There are many challenges in tokenization which depends on the type of language [21].

The tokenization process of the system uses Myanmar 3 based tokenizer. It is based on Myanmar writing format, start with one of consonants, vowel, number, special character. For example, for the input Myanmar sentence “နွေဥတုသည်ပူပြင်းသည်”, the tokenization process of the system tokenizes the sentence as နွေ/ဥ/တု/သည်/ပူ/ပြင်း/သည်.

5.1.2 Segmentation

In Myanmar script, sentences are clearly delimited by a sentence boundary marker, but words are not always delimited by spaces. Therefore, segmenting Myanmar sentences into word is a challenging task. Segmentation process constructs the Myanmar words form syllables. In the segmentation process of the system, input Myanmar sentences are separated into words by using longest matching approach with the aid of Myanmar to English bilingual lexicon. As an example, for the input sentence, “နွေဥတုသည်ပူပြင်းသည်”, the process will generate နွေဥတု/သည်/ပူပြင်း/သည်.

5.1.3 Part of Speech (POS) Tagging

In Myanmar, there are eight parts of speech: noun, pronoun, verb, adverb, adjective, preposition, conjunction, and particle as described in Table 1. Part of speech tagging process of the system assigns each word of the sentence with these tags [19]. The part of speech tagging process of the system will tag the segmented sentence နွေဥတု/သည်/ပူပြင်း/သည် as follows:

- နွေဥတု-common countable singular
- သည်-subject preposition

- ပူဖြင်း-verb
- သည်-verb preposition

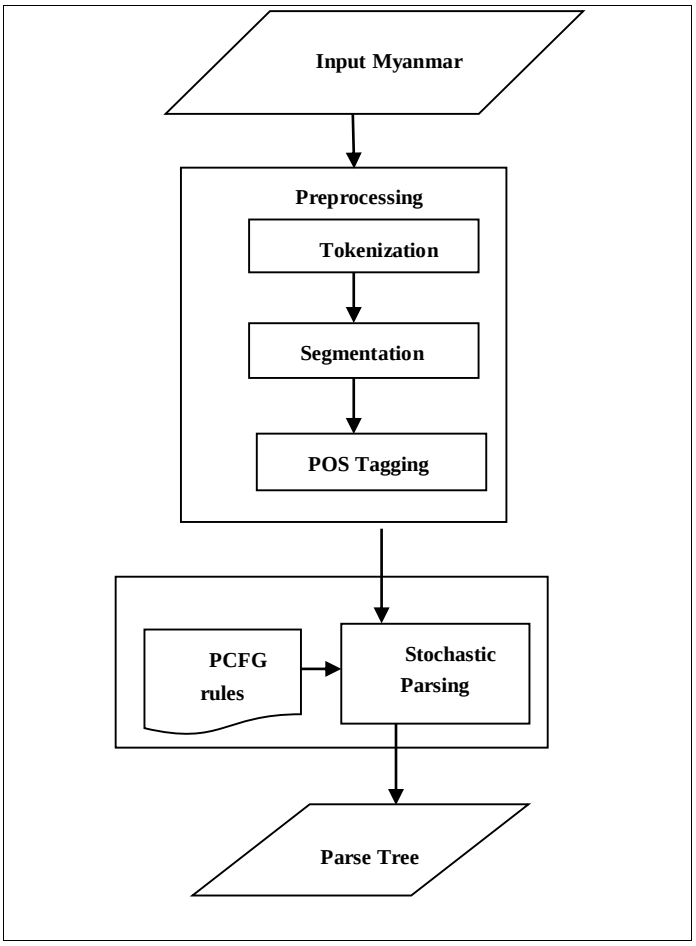

Figure.1 Proposed Stochastic Parsing System

5.2 Stochastic Parsing

Parsing is an analysis of a sentence which first identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and then links them to higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc.). It is a technique widely used in natural language processing. It is similar to the concept of lexical analysis for computer languages. Under the name of the Shallow Structure Hypothesis, it is also used as an explanation for why second language learners often fail to parse complex sentences correctly.

5.2.1 Context Free Grammar

A context free grammar (CFG) can be defined as the following 4 tuple:
G= (N, Σ,R,S), where

1. N is a set of nonterminal symbols
2. Σ is a set of terminal symbols.
3. R is a set of rules or productions in the form of A→β , where A ∈N and β is an ordered list of symbols drawn from N ∪Σ.
4. S is the starting symbol

Sentences can be generated from a CFG in the following derivation process: Starting with S, rewrite a nonterminal A by replacing it with the right-hand side of a rule with A on the left-hand side. Repeat this rewriting process until we end up with a string of terminals. One view of parsing is to recover this derivational process for a target output sentence. Because, natural language is highly ambiguous, many CFG parse can result from a single sentence [5]. For Myanmar Language, these Context Free Grammar rules are defined with the Part of Speech (POS) sequences of Myanmar language [12]. The example CFG rules are presented in Figure 2.

NP→Noun POPOS NCCS PAIDNUM
NP→PP Noun PAIDNUM
NP→PP NP
NP→ Noun POPOS NP
PP→NounPOPOS
NP→NounPAIDNUM
Noun→ NCCS
Noun→NCPS

Figure.2 Example CFG Rules

5.2.2 Stochastic Context Free Grammar

Probabilistic context free grammar is an extension of CFGs in which each rule is associated with a probability $p \in [0, 1]$. In a consistent PCFG, the probabilities of all the rules with the same nonterminal on the left-hand side from a probability distribution are sum to one.

NP→Noun POPOS NCCS
PAIDNUM ,0.053
NP→PP Noun PAIDNUM ,0.043
NP→PP NP, 0.02
NP→ Noun POPOS NP, 0.05
PP→ Noun POPOS ,
NP→ Noun PAIDNUM

Noun→ NCCS , 0.083

Figure.3 Example PCFG Rules

The main utility of PCFG is to select the best parse for a sentence among multiple parse, according to its probability model. The probability of a parse tree in a PCFG is the product of the probabilities of each of the rules in the parse tree [5]. Some of these PCFG rules are described in Figure 3.

6 Conclusions

Chunking or shallow parsing segments a sentence into a sequence of syntactic constituents. In this paper, stochastic parsing system for Myanmar language is presented. In this system, Context Free Grammar and Stochastic Context Free Grammar (SCFG) will be used for parsing of Myanmar sentence. There are two types of Myanmar Language, written language and spoken Language. The system presented in this paper is based on the written form Myanmar Language. For the generation of Grammar rules (i.e. CFG and PCFG), example sentences from Myanmar Grammar book were used as training sentences. This CFG and SCFG will be applied in bottom up parser and then the parse trees of Myanmar sentences will be generated and the accuracy of two grammars, CFG and PCFG, will be compared. Parsing of a sentence can help many other NLP applications such as text summarization, machine translation, phrase chunking and information retrieval systems.

References

1. D.L.Cing, T.M. Htwe, A Probabilistic Model for Joint Myanmar Morphological Segmentation and Part of speech Tagging, International Conference on Computer Applications, February 2017, Yangon, Myanmar.

2. E.Keryszig, Advanced Engineering Mathematics, Seventh Edition.
3. F.Feng and W.Bruce Croft, “Probabilistic Techniques for Phrase Extraction”, Journal of Information Processing and Management: an International Journal, Volume 37, 2001.
4. F.Oliveraet at, “Systematic Noun Phrase Chunking by Parsing Constraint Synchronous Grammar in application to Portuguese Chinese Machine Translation”, Proceedings of the sixth International Conference on Information Technology Applications, 2009.
5. J.C.K. Cheung”, “Parsing German Topological Fields with Probabilistic Context-Free Grammars”, M.Sc. Thesis, Graduate Department of Computer Science, University of Toronto.
6. J.Okell, A.allott, “Burmese/Myanmar Dictionary of Grammatical Forms”, Cruzan Press, 2001.
7. K.K. Batra and GS Lehal, “Rule Based Machine Translation of Noun Phrase from Punjabi to English”, International Journal of Computer Sciences Issues, 2010.

8. K.Karoo, G.Katkar, “Analysis of Probabilistic Parsing in NLP”, International Journal of Innovative Research in Computer and Communication Engineering. Volume 4, Issue 10, October 2016.
9. M. P. Aung, K. T. Lynn, “Construction of Finite State Machine for Myanmar Noun Phrase”, Proceedings of MJIIT_JUC Joint International Symposium, 2013, Hiratsuka, Japan, (2013).
10. M. P. Aung, A. L. Moe, “New Phrase Chunking Algorithm for Myanmar Natural Language Processing”, International Journal of Applied Mechanics and Materials, volume 695, pp 548-552, (2015). 548-552, (2015).
11. M.P.Aung, “Constructing Myanmar Phrase Structure Grammar for Myanmar Noun Phrase Extraction”, the seventh International Conference on Science and Engineering, December 2016, Yangon, Myanmar.
12. M.P.Aunget. al, “Stochastic Context Free Grammar for Statistical Parsing of Myanmar”, International Conference on Computer Applications, February 2018, Yangon, Myanmar.
13. M.T.Win et al, “Burmese Phrase Segmentation”, Proceedings of Conference on Human Language Technology for Development, Alexandria, Egypt, May 2011, pp27-33.
14. P. M. Hopple, “The Structure of Nominalization in Burmese”.
15. P.M.Nugue, “An Introduction to Language Processing with Pearl and Prolog”.
16. S. P. Soe, Dr. “Aspects of Burmese Language”, Department of Myanmar, UDE university of Yangon, (2010).
17. S.S Lwin, Yuzana, K. T. Nwet, Framework of the Extractive Myanmar Text Summarization with Naïve Bayes Classifier, International Conference on Computer Applications, February 2017, Yangon, Myanmar.
18. S.T.Y.Myint, M.M.Khin, “Lexicon Based Word Segmentation and Part of Speech Tagging for Myanmar Text”, International Journal of Computational Linguistics and Natural Language Processing, Volume2, Issue6, June 2013, pp 394-403.
19. S.Myint, M.M.Khin, “Lexicon Based Work Segmentation and Part of Speech Tagging for Written Myanmar Text”, International Journal of Computational Linguistics and Natural Language Processing, Volume 2, Issue 6, June 2013, pp 396-493.
20. Y.M.S.Yin, Yuzana, Khin Mar Soe, Identifying Myanmar Phrases Using Conditional Random Field.
21. www.language.worldofcomputing.net/category/tokenization.