Finite-state Approach based Myanmar Morphological Analysis

Tin Myo Latt, Aye Thida
University of Computer Studies, Mandalay
tinmyolatt.tml@gmail.com, ayethida.royal@gmail.com

Abstract

Morphological analysis (MA) is needed in any Natural Language Processing (NLP) Application. It means taking a word as input and identifying the stem and affix. MA provides information about a word's semantics and the syntactic role which plays in a sentence. This paper presents the development of a Myanmar morphological analysis; morphological processes prevalent in Myanmar language are explored. We consider Myanmar MA for noun, verb, adjective and adverb. Myanmar morphology reveals the three types. They are inflectional morphology, derivational morphology and compounding. In this work, finite state automaton (FSA) is used to model Myanmar morphology which contains a monolingual lexicon. MA will be to apply as a portion of the Grammar Checker for detecting grammatical errors in Myanmar texts. The proposed framework of this paper is to describe Myanmar morphological analysis.

Keywords: *Morphology, finite-state automaton, Morphological Analysis*

1. Introduction

Natural Language Processing (NLP) is a set of computational techniques for analyzing and representing text in Natural Language (NL) with linguistic analysis for achieving human-like language processing for a range of tasks or applications. It deals with interactions between computer and human (natural) languages. Morphology is the field of the linguistics that studies the internal structure of the words.

MA allows us to reduce the size of the dictionary (lexicon), but we need a list of exception for every morphological rule we invent. MA refers to the computational processes which provide structural information about surface words in a language. The MA of a word is the investigation through the identification and study of morphemes, often defined as the smallest linguistic pieces with a grammatical function. So computationally the MA of a word constitutes taking a word form as input and producing the structure of the word by showing the lexical category of the constituent morphemes.

This paper is organized as follows: Section 2 discusses the MA used for initial analysis of the input text as well as previous work that has been done in the area for languages. Section 3 describes the nature of Myanmar grammatical categories. Section 4 states Myanmar morphology applying the FSA. Section 5 shows the building of Finite-state Automaton for Words and section 6 contains final conclusions.

2. Related work

Ksh. Krishna B. Singha et.al [5] proposed a constrained finite-state model to represent the morphotactic rule of Manipuri adjective word forms. There was no adjective word category in Manipuri language. By rule this category was derived from verb roots with the help of some selected affixes applicable only to verb roots. Finite-state machine was used to describe the concatenation rules and corresponding nondeterministic and deterministic automaton were developed for ease of computerization. A root lexicon of verb category words was used along with an affix dictionary in a database. The system was capable to analyze and recognize a certain word as adjective by observing the morpheme concatenation rule defined with the help of finite-state networks.

Soe Lai Phyue et. al [7] presented the morphological processor (analyzer and generator), Morphocon, to support the inflectional verbal and colloquial cases for knowledge resourcesby using the rule-and-feature based model of Myanmar inflectional morphology. By supporting with Morphocon in Myanmar Language Resources, it could reduce the time and storage consumption. The evaluation of the correctness of Morphocon yields the satisfactory result because precision,

recall and f-measure are nearly and over 95% in both morphological analyzer and generator.

3. Morphology

Morphology is the study of the way words are built up from smaller meaning-bearing units, morphemes. Morphemes are either free or bound forms, with the free forms corresponding to word level units and the bound forms to a closed class of grammatical affixes. For example, the word မြစ် (river) consists of a single morpheme (the morpheme မြစ်) while the word ကြောင်များ (cats) consists of two: the morpheme ကြောင် (cat) and the morpheme များ (-s). [2]

Morphemes are divided into two types, open class and closed class. Open class items belong to categories/types to which new members may be freely added. Closed class items on the other hand belong to categories/types to which new members cannot be added.

There are many ways to combine morphemes to create words. In this paper presents three of these methods which are common inflection, derivation and compounding for Myanmar morphology.

3.1 Inflectional Morphology

Inflectional is the combination of a word stem with a grammatical morpheme, usually syntactic resulting in a word of the same class as the original stem, and usually filling some syntactic function like agreement. Myanmar has a relatively simple inflectional system which contains noun, verb and adjective, not adverb.

3.2 Derivational Morphology

Derivation is the combination of a word stem with a grammatical morpheme, usually resulting in a word of a different class, often with a meaning hard to predict exactly. For example, the verb စား can take the derivational suffix ခြင်း to produce the noun စားခြင်း.

It is not at all unusual for derivational affixes to change verbs into nouns or adjectives, adjectives into nouns or verbs, that sort of thing. Derivational affixation can change category.

3.3 Compounding morphology

Compounding is the combination of multiple words stems together. For example, ပဲပြုတ်/pe: bjou`/ (boiled pea), နေ့စဉ်မှတ်တမ်း/nei. zin hma` tan: / (diary), မျက်နှာစုံညီစည်းဝေးပွဲ/mje`hna soun nji si wei: pwe: / (plenary Meeting).

4. Finite State Morphological Parsing

A Myanmar word will divide into smaller subdivisions. For example, if a word is given ကစားသည် (play) to the morphological parser it will generate the output ကစား + V and သည် + PresentTense. ကစား is the root morpheme and သည် is postposition of verb (PresentTense), are morphological features. These features specify the additional information about the stem. In order to build a morphological parser we need at least the following: (1) Lexicon (2) Morphotatics (3) Orthographic Rules. [2]

In this paper will express Lexicon and Morphotatics for Myanmar morphological analysis. Although English language requires orthographic rules such as consonant doubling rule, E insertion and E deletion rule etc…, there is no need in Myanmar language.

4.1. Lexicon

The list of stems and affixes, together with basic information about them (whether a stem is a Noun stem or a Verb stem, etc.). Every lexicon is of a certain class. The following example:

Morpheme\1:
ကစား/gaza (play)
Class: Verb_Stem or Root
Feature: Parts of Speech = Verb
Morpheme2:
ခြင်း/chin (particle for noun phrase change)
Class: Noun_ Suffix
Feature: Parts of Speech = Particle

All the lexicons in a certain class are stored in a FSA. Myanmar morphological analysis will need the lexicon which contains the stems and affixes.

4.2. Morphotatics

Myanmar morphology is rich and complex. Morphotactics represent the ordering restrictions in place on the ordering of morphemes. Morphotactics can be concatenative, with morphemes either prefixed or suffixed to stems. A basic morphotactic fact about affixes is where they attach with respect to the stem.

Prefix + Stem + Suffix

An affix is either a prefix or a suffix; Plural - များ၊ တို့၊တွေ is a suffix, အ- is a prefix. Myanmar morphological analysis applies the building of FSA.

5. Building of Finite-state Automaton for Words

The objective of the FSA can use Myanmar morphology to solve the problem of determining whether an input string makes up a legitimate Myanmar word or not in the language. Given an input string, an FSA will either accept or reject the input. An FSA can use set of symbol for its alphabet, including words. FSA using all possible affixes is built.

An FSA defines a language to be

- A set of strings over some alphabet $\sum$
- A set of states Q
- A designated start state q_0 ($q_0 \epsilon Q$)
- A set of accepting final statesq_j ($q_j \subset Q$)
- Edges: given current state q_i and input x $\in \sum$, gives new state q_j

5.1. Inflection of Noun

Myanmar nouns are regular nouns and irregular nouns. They have three kinds of inflection and an affix marks plural. Nouns in Myanmar are pluralized by suffixing the particle တွေ [twe] in colloquial Burmese or များ [myar] in formal Burmese. The particle တို့ [tou] which indicates a group of people or things is also suffixed to the modified noun.

The numbered circles (nodes) represent states and the labeled arcs represent transitions from one state to another. Here the start state is the circle numbered with q_0. The double circle denotes the final (accepting) state. The labels with each arc suggest that a transition is possible only when the labeled string is matched with the input text.

The FSA in Figure 1 assumes that the lexicon includes regular nouns and irregular noun that take the regular - များ၊ တို့၊တွေ plural.

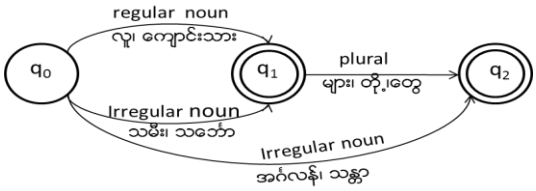

Figure 1. An FSA for a fragment of noun inflectional morphology

5.2. Inflection of Verb

Verbs have three tenses for Myanmar language. There are the present tense, followed by the past tense, and future tense. A verbal postposition used to express the same as present and past tense of the verb is called a verbal postposition of present tense. It is သည်/thi/ (word indicating the verb ending a sentence), ၏/i/ (word placed at the end of an affirmative sentence), ပြီ/pji/ (word following a verb indicating that an action is taking place or has already taken place). The future tense of the verb is called a verbal postposition of future tense. It is မည်/ mji/ (shall or will), လိမ့်မည်/lein mji/ (shall or will), အံ့ /an/ (shall or will), လတ္တံ့/latan/ (shall or will) [3].

The word ခဲ့/khe is particle suffixed of verbs to emphasize definitiveness of action or condition. It is not expressed as suffixed the past tense in MLC, 2006 [3, page-253]. So it can not contain in the building of the FSA in figure 2.

The FSA in Figure 2 shows the lexicon includes verb stem plus three more suffix (present, past and future tense). Particle is between stem and suffix. It has five states. If state q_0 is the start state and input word is verb in lexicon then changes the state q_1, reading the next word is particle, change to the next state q_2, continuous reading from the next input word is tense; change to the state q_3 is final state. Another path, the state q_1 and input is word of tense that change to q_3 is final state.

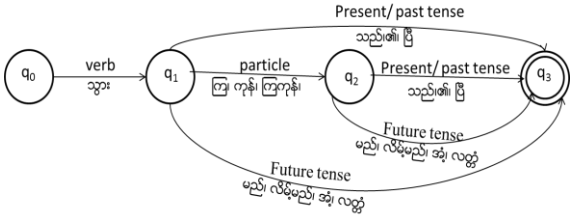

Figure 2. An FSA for a fragment of verb inflectional morphology

5.3. Inflection of Adjective

Adjectives can be divided into three stages: the normal stage (positive degree), the superior stage (comparative degree) and the most superior stage (superlative degree). The normal stage is the base form of adjective that precedes the head

nominal and is marked with the particle သော၊ သည့်၊ မည့်. The comparative degree is expressed in Myanmar by "သာ၍/ (more) or ပို၍/pou jwei./ (more)" while in English the comparative degree is marked by adding the prefix "more" before an adjective or by the suffix "er" after an adjective. In Myanmar, the superlative degree is formed by prefixing "အ/a./" and affixing "ဆုံး/hsoun:/ (most)" to the adjective. In Figure 3, the start state is q_0 and the final state is q_5.

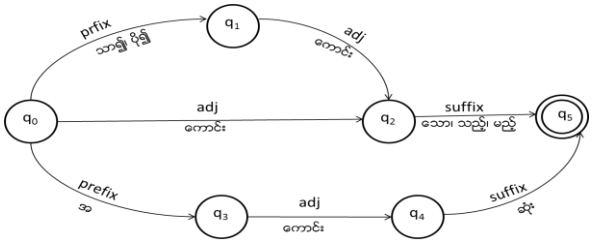

Figure 3. An FSA for a fragment of adjective (degree) inflectional morphology

5.4. Derivation of Noun

A noun formed by using a particle before or after an attributive word is called an attributive noun. The particles used to join together with an attributive word to form an attributive noun are "အ/a/, မှု/mu./and ခြင်း/chin:/". For example, ကောင်းမှု/kaun: mu./ (good deed), လှခြင်း/hla. kjin:/(beauty), ထူးခြားချက်/htu: kja: kje./ (being distinctive).

In Figure 4, the start state q_0, the input is the adjective or verb then changes to state q_1. And then state q_1, the next input is suffix then changes to the final state q_2.

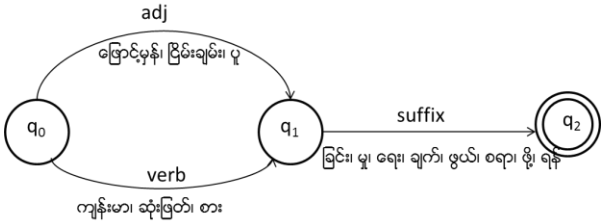

Figure 4. An FSA for a fragment of noun derivational morphology

5.5 Derivation of Verb

An adjective and a verbal postposition can be combined to form a derivation of verb in Myanmar. In Figure 5 as shown the FSA will recognize the adjectives followed by tense or particle. If the next input is particle then follow by tense.

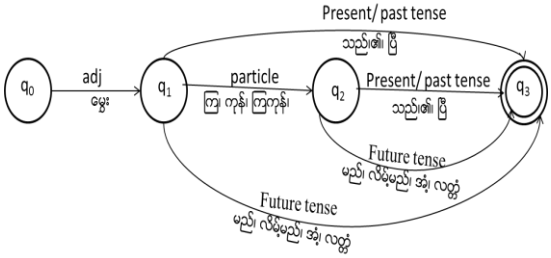

Figure 5. An FSA for a fragment of verb derivational morphology

5.6. Derivation of Adjective

In Myanmar, the use of adjective-forming particles as 'သော /tho:/ (adjective), သည့်/thi./ (which, that) and မည့်/mji./ (really)' together with verbs is found in the way English verbal adjectives do.

In Figure 6, the FSA will recognize the part of speech of input word is the verbs; the starting state is q_0, and then changes to state q_1. And then state q_1, the next input is suffix then changes to the final state q_2.

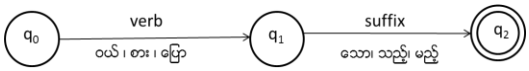

Figure 6. An FSA for a fragment of Myanmar adjective derivational morphology

5.7. Derivation of Adverb

Adverbs make the sense of the sentence more profound by combining word classes in terms of structure apart from meaning. In most cases, the term adverb is not a major one in the structure of the sentence.

Adverbs in terms of structure are reduplicated adverbs, affixed adverbs, rhyming adverbs. Some adverbs can express in this paper, not all of adverbs.

A particle-suffixing adverb in Myanmar is an adverb formed by affixing the particle 'စွာ/ swa/ (-ly)' after a verb or an adjective.An adverb of manner in Myanmar is a word used to modify a verb expressing how someone behaves or something is done [1]. For example, ရိုသေစွာ /jou thei swa/ (respectfully), လျင်မြန်စွာ/hlin mjan swa/ (quickly), ခင်မင်စွာ/khin min swa/ (affectionately), ချိုသာစွာ /chou tha swa/ (sweet and approving).

In Figure 7, the FSA will recognize the verbs or adjectives, start state is q_0, and part of speech of input word is the verb or adjective then changes to state q_1. And then state q_1, the next input is suffix then changes to the final state q_2 (accept), otherwise reject.

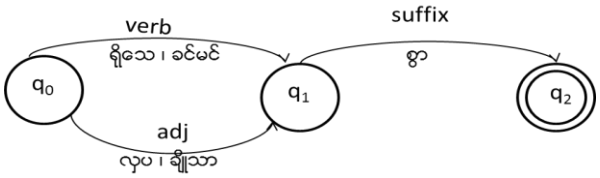

Figure 7. An FSA for a fragment of adverbs (particle-suffixed) derivational morphology

A particle-affixing adverb (mid and end) in Myanmar is an adverb formed by affixing such particles as 'ချည်-ချည် /chi chi/, လိုက်-လိုက် / lai`/' after a verb or an adjective, or in the middle of them [1]. For example, ဝင်ချည်ထွက်ချည်-/win chi htwe` chi/ (coming in and out alternately), ပူလိုက်အေးလိုက် /pu lai` ei: lai`/ (being hot and cold alternately. Then particle-affixing such as မိ-ရာ and some verbs such as တွေး၊ ငေး၊ ထင်၊ ပြော combine and the combinations are used as adverbs.

In Figure 8, the FSA will recognize the verbs, adjectives, reduplication verb. The start state is q_0; input word is verb in lexicon then changes to state q_1. The start state is q_0; input word is adjectives then changes to state q_2. The start state is q_0; input word is adjectives then changes to state q_2. The start state is q_0; input word is reduplication verb then changes to state q_3. And then state q_1, the next input word is infix and suffix then changes to the final state q_4 (accept), otherwise reject.

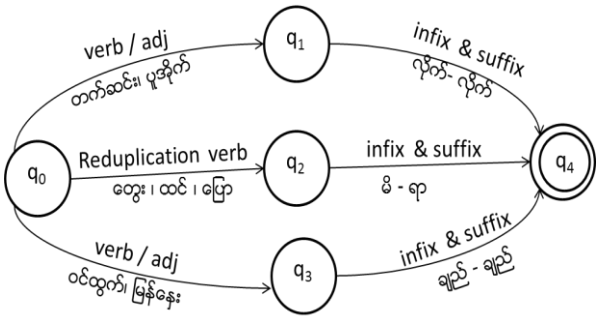

Figure 8. An FSA for a fragment of adverbs (particle-infixed and suffixed) derivational morphology

5.8. Compounding of Noun

Compound noun refers to a noun which joins a noun, a verb, a pronoun, an adjective and an adverb together accordingly without putting prepositions, particles and conjunctions between them [1].

The FSA makes a choice from the starting state q_0, going either to q_1 and q_2, which are the new states corresponding to old state q_0 and input noun or verb. If the FSA selects to q_1, part of speech of input word is noun or adjective, the new state is q_3 or q_5 which are final state. It continues to operate in this processing and there may be many choices. The final states have one or more states which are state q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}, q_{12}.

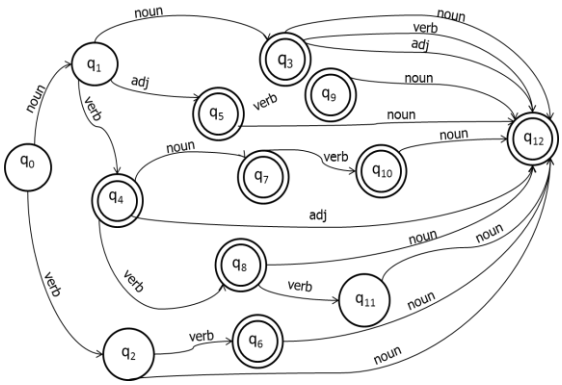

Figure 9. An FSA for a fragment of noun compounding morphology

Table 1. Myanmar compounding morphology for some nouns

No	POS				Compound Noun
1	ပင်လယ် (noun)	ရေ (noun)	ငန် (adj)	-	ပင်လယ်ရေငန်
2	စိန် (noun)	ရင် (noun)	ထိုး(verb)	-	စိန်ရင်ထိုး
3	မျက်နှာ (noun)	စုံ (verb)	ညီ (verb)	-	မျက်နှာစုံညီ
4	ရုပ် (noun)	မြင် (verb)	သံ (noun)	ကြား (verb)	ရုပ်မြင်သံကြား
5	ပေါင်း (verb)	ကူး (verb)	တံတား (noun)	-	ပေါင်းကူးတံတား
6	ရုံး (noun)	သုံး (verb)	ဘာသာ (noun)	စကား (noun)	ရုံးသုံးဘာသာစကား

5.9. Compounding of Verb

Compound verbs of English are verbs affixed by a verbal postposition on the formation of two words. For example, နှုတ်ခွန်းဆက်သ/hnou` khun: hse` tha. thi/ (greet).

The FSA can check part of speech of the input word whether accept or reject. If it corrects the compound verb then it will be accepting.

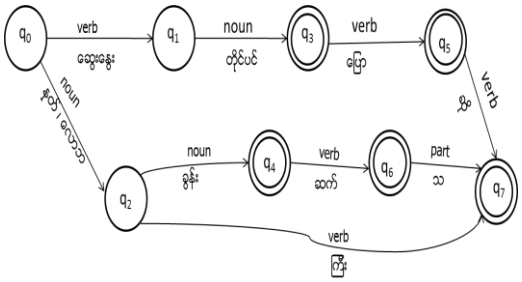

Figure 10. An FSA for a fragment of verb compounding morphology

5.10. Compounding of Adjective

A compound adjective in Myanmar is an adjective, consisting of at least an adjective, formed by the adjective and a noun or another adjective [1]. For example, ရုပ်ဖြောင့်/jou` hpjaun./ (handsome), သေးသွယ်/thei: dhwe/(slim), သတ္တိပြောင်/tha` ti. pjaun / (bold).

In Figure 11, FSA starts in state q_0, an input of adjective of words will change to state q_1, and an input of adjective will choose either state q_2 or q_3. If it will select state q_4 for the checking of two consecutive adjectives of words and this state q_4 is finial state (accept). But it will select state q_2 will continue another states reaching to final state. This means that the input words are valid for compound adjective. If it will not reach to state q_4 (reject state) then the input words are invalid.

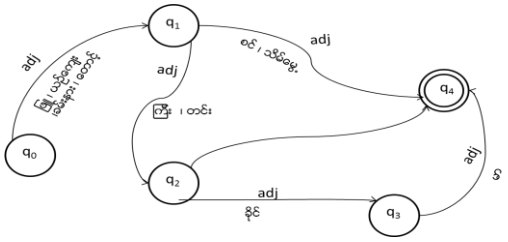

Figure 11. An FSA for a fragment of adjective compounding morphology

5.11. Compounding of Adverb

Compound adverbs of Myanmar are formed by adding together_ noun and noun or noun and adjective or verb and verb [1]. For example, ပိုက်စိပ်တိုက်/pai` sei` tai`/ (searching closely), ကောင်းကောင်း /kaun: kaun:/ (well)

In Myanmar, adverbs are formed two adjectives or verb can be joined together to form an adverb. Such kind of adverb is called a double adverb having one word မြန်မြန် (quickly), ခင်ခင်မင်မင် (with friendliness), ချိုချိုသာသာ (sweetly). Three adjectives can also be joined together to form an adverb ချစ်ချစ်တောက် (blazingly, blisteringly, feverishly), စိုထိုင်းထိုင်း (be humid, be damp).This ချစ်ချစ်တောက် is divided into two morphemes, ချစ်ချစ် and တောက်. The morpheme ချစ် (burnt) is adjective.[3]

In Figure 12, FSA starts in state q_0, an input of adjective of words will choose the state q_1, q_2, q_3, q_4, q_5 and an input of adjective or verb. If it has selected one state then it moves to finite state q_5. Otherwise it will reach rejecting state.

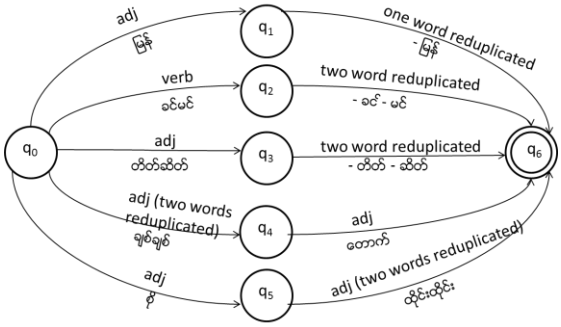

Figure 12. An FSA for a fragment of adverbs (reduplicated) compounding morphology

6. Conclusion

Morphological analysis is very important and basic applications of Natural Language Processing. Morphological analysis needs for Myanmar language because Myanmar language is morphologically rich and agglutinative language. This paper describes the framework of morphological analysis based on finite-state automaton approach for Myanmar word class. It reveals the framework but not implementation. Most of this works are focus on analysis of noun, verb, adjective and adverb.

References

[1] Aung Zin Minn,*A Comparative Study of the Two Grammatical Systems of Written English & Myanmar and Its Significance to Learning English as a foreign language*, Department of English, University of Mandalay, Myanmar, May, 2009

[2] Daniel Jurafsky & James H. Martin, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Copyright c 2006, All rights reserved. Draft of June 25, 2007.

[3] Department of the Myanmar Language Commission, Myanmar-English Dictionary, Ministry of Education, Myanmar, 2006.

[4] http://en.wikipedia.org/wiki/ Burmese_ Language

[5] Ksh. Krishna B. Singha et. al, "Morphotactics of Manipuri Adjectives: A Finite State Approach", I.J. Information Technology and Computer Science, 2013, 09, 94-100

[6] Paulette M. Hopple, *The Structure of Nominalization in Burmese*, SIL International 2011

[7] Soe Lai Phyue et. al, "Morphological Processor for Inflectional Case of Multipurpose Lexico-Conceptual KnowledgeResource", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 1, Issue 7, September 2012wledge Resource,

[8] Thang Khan Dim et. al, A Contrastive Study of Adverbs of Manner in German and Myanmar, The Government of The Republic of the Union of Myanmar Ministry of Education, Universities Research , Journal 2012, Vol. 5, No. 7