Myanmar Word Segmentation

Hla Hla Htay, Kavi Narayana Murthy

Department of Computer and Information Sciences
University of Hyderabad
email:knmuh@yahoo.com ,hla_hla_htay@yahoo.co.uk

Abstract

Myanmar is written without necessarily pausing between words with spaces. It is therefore non-trivial to segment sentences into words. Word tokenizing plays a vital role in most Natural Language Processing applications such as Summarization, Information Retrieval, Machine Translation, Text Categorization, and so on. In this paper we describe our initial attempts at word segmentation in Myanmar. We exploit lists of stop words and n-grams at syllable level to build a word hypothesizer. We have achieved about 65% accuracy in word hypothesis. We have built a manually checked word list of nearly 100,000 words in Myanmar.

1. Introduction

Myanmar language is similar to other Asian languages including Indian languages, Chinese, Japanese and Thai. According to history, Myanmar script has originated from Brahmi script which flourished in India from about 500 B.C. to over 300 A.D [11].

In Myanmar script, sentences are clearly delimited by a sentence boundary marker but words are not always delimited by spaces. Although there is a general tendency to insert spaces between phrases, inserting spaces is more of a convenience rather than a rule. Spaces may sometimes be inserted between words and even between a root word and the associated post-position. In fact in the past spaces were rarely used. Segmenting sentences into words is therefore a challenging task. We show below an example of a sentence being segmented into words:

လေကောင်းလေသန့်သည်ကျန်းမာရေးအတွက်ကောင်းသည်။

လေကောင်းလေသန့်.	ကျန်းမာရေး	ကောင်း
le kaung: le than.	kyan: ma ye:	kaung:
fresh air	health	good
N	N	Adj

In this paper we describe our initial attempts at segmenting Myanmar sentences into words. After a brief discussion of the corpus collection and pre-processing phases, we describe two approaches to segmentation, one based on a list of stop words and the other using n-grams of syllables.

Since dictionaries and other lexical resources are not yet widely available in electronic form for Myanmar language, we will not be able to exploit stored word lists. On the other hand, our work here can in fact lead to preparation of electronic word lists, dictionaries and other lexical resources. We now have a collection of nearly 100,000 inflected words. These words have been manually checked. Further work is on.

Development of electronic dictionaries will facilitate Natural Language Processing tasks such as Spell Checking, Machine Translation, Automatic Text summarization, Information Extraction, Automatic Text Categorization, and Information Retrieval and so on [2]. Further, we will be able to develop thesauri, word-nets, annotated corpora, morphological analyzers etc.

2. Preprocessing

Development of lexical resources is a very tedious and time consuming task and purely manual approaches are too slow. We have downloaded Myanmar texts from various websites including news sites and on-line magazines. As of now, our corpus includes more than 75000

sentences. The downloaded corpora need to be cleaned up to remove hypertext markup etc. We have developed the necessary scripts in Perl. Also, different sites use different font formats and character encoding standards are not yet widely followed. We have mapped these various formats into the standard Win Innwa font format. We have stored the cleaned up texts in ASCII format. This will enable processing in environments where Unicode is not yet supported. It is easy to switch to Unicode where required.

3. Stop Word Removal

Preliminary studies suggested that Myanmar sentences can be tokenized by eliminating stop words. Hopple [6] also notices that particles ending phrases can be removed to recognize words in a sentence. Stop words are defined as non-information-bearing words. They form closed classes and hence can be listed. Stop words include prepositions/post-positions, conjunctions, particles, inflections etc. These words appear so frequently that their usefulness is limited. In *Information Retrieval*, for example, search engines ignore stop words at the time of searching a key phrase. In *Information Extraction and Text Summarization* also, stop words are pushed aside and treated as irrelevant information, in order to extract the most relevant and important information.

We have collected stop words by analyzing official newspapers, Myanmar grammar text books and CD versions of *English-English-Myanmar (Student's Dictionary)* [7], *English-Myanmar Dictionary* [8], and *The Khit Thit English-Myanmar dictionary* [4]. We have also looked at stop word lists in English [13] and mapped them to equivalent stop words in Myanmar. See Table 4. As of now, our stop words list contains about 1216 entries. Stop words can be prefixes of other stop words leading to ambiguities. Usually, the longest matching stop word is the right choice.

Below are some example sentences where we show this longest matching stop word recognition leading to correct word segmentation. In each case, we show the original sentence, segmented words, roman transliteration and gloss in English. In our work here we have removed spaces before processing since spaces are not dependable.

1. ဝိုင်းဝန်းချီးကျူးခံရသဖြင့်သူအနေခက်ရသည်

ဝိုင်းဝန်းချီးကျူးခံရ	အနေခက်
waing: win: chi: kyu: khan ya	a nay khak
(received compliments)	(abashed)
Vpp	Vpast

2. ကျောင်းအုပ်ဆရာကြီးသည်အကြမ်းဖက်မှုကိုစက်ဆုပ်သည်

ကျောင်းအုပ်ဆရာကြီး	အကြမ်းဖက်မှု	စက်ဆုပ်
kyaung: aop hsa ya kyi:	a kyan: phak mhu	sak sop
The headmaster	violence	abhors
Nsubj	Nobj	Vpresent

4. Collecting words using Text:: Ngrams

As we keep analyzing texts, we can identify some words that can appear independently without combining with other words or suffixes. We build a list of such valid words and we keep adding new valid words as we progress through our segmentation process, gradually developing larger and larger lists of valid words. This list of known words can be made use of for hypothesizing candidate words as we go along.

Myanmar language uses a syllabic writing system [2] unlike English and many other western languages which use an alphabetic writing system. Interestingly, almost every syllable has a meaning in Myanmar language. This can also be seen from the work of Hopple [6].

Myanmar Natural Language Processing Group has listed 1894 syllables that can appear in Myanmar texts. We have observed that there are some more syllables, especially in foreign words including Pali and Sanskrit words which are widely used in Myanmar. We have collected other possible syllables using Myanmar-English dictionary [5]. Now we have over 2000 syllables in our list.

We have developed scripts in Perl to syllabify words using our list of syllables and then generate n-gram statistics using Text::Ngrams

which is developed by Vlado Keselj [12]. It is a very fast program and it took only 5 minutes on a desktop PC in order to process 3.5M bytes of Myanmar text file. We have used *"—type=word"* option treating syllables as words. We had to modify this program a bit since Myanmar uses zero (as "(o) wa " letter) and the other special characters (",", "<", ">", ".", "&", "[", "]" etc.) which were being ignored in the original Text::Ngrams software.

We collect all possible n-grams of syllables upto 5-grams. Table 1 shows some words which are collected through n-gram analysis. Almost all monograms are meaningful words. Many bi-grams are also valid words and as we move towards longer n-grams, we generally get less and less number of valid words. See Table 2. Further, frequency of occurrence of these n-grams is a useful clue. Techniques such as mutual information and maximum entropy can be used to hypothesize possible words. Manual checking is essential to finally choose valid words. Words already recognized can be used to hypothesize more words. For example, if one part of a sequence is an already known word, the other part is a candidate word. Length statistics will be a useful hint and many researchers have used longest string matching [9, 10]. Here we have used a combination of some of these ideas to prepare an initial list of valid words. See Table 5. So far nearly 100,000 valid words have been identified. With this in background, it is now possible to apply machine learning techniques to develop an intelligent word hypothesizer. Further work is on.

There are lots of valid words which are not described in published dictionaries. The entries of words in the Myanmar-English dictionary which is produced by the Department of the Myanmar Language Commission are mainly words of the common Myanmar vocabulary. Most of the compound words have been omitted in the dictionary [5]. This can be seen in the preface and guide to the dictionary of the Myanmar-English dictionary produced by Department of the Myanmar Language Commission, Ministry of Education. 4-syllables words like "ထူးထူးဆန်းဆန်း" (strange), "ထူးထူးကဲကဲ" (outstanding) and "ထူးထူးခြားခြား" (different) are not listed in dictionary although we usually use those words in every day life. Statistical construction of machine readable dictionaries has many advantages. New words which appear from time to time such as internet, names of medicines, can also be detected. Compounds words also can also be seen.

Table 1. Example of collected Ngrams

Bigram(two syllables)	3-grams(three syllables)	4-grams(four syllables)
ဖန်ထည်	ဝါးခနဲ	ငယ်ငယ်ကြီးကြီး
ဖန်တုံး	ဝုန်းခနဲ	ခါးခါးသီးသီး
ဖန်တီး	ဝေါခနဲ	တောင့်တောင့်တင်းတင်း
ဖန်လာ	ဒေါင်ခနဲ	နှစ်နှစ်သက်သက်
ဖန်အိမ်	ဗုန်းခနဲ	နှစ်နှစ်ကာကာ
ဖန်ဆင်း	ဗွေခနဲ	ပျစ်ပျစ်နှစ်နှစ်
ဖန်သား	ဗွေ့ခနဲ	ပုံ့ပုံ့စားစား
ဖန်ခုန်	ဆောင့်ခနဲ	စွန့်စွန့်စားစား
ဖန်ခွက်	မှုးခနဲ	စဉ်းစဉ်းစားစား
ကန်စောင်း	ထောင်းခနဲ	များများစားစား
ကန်ရေ	ဖောက်ခနဲ	သက်သက်သာသာ
ကီလို	ဘုတ်ခနဲ	ကြီးကြီးမားမား
ကန်ဇွန်း	ချွတ်ခနဲ	ဆန်းဆန်းကျယ်ကျယ်
ကန်တော့	နှင့်ခနဲ	ဆန်းဆန်းပြားပြား
ကင်းငြိမ်း	ပြက်ခနဲ	ကျယ်ကျယ်ပြောပြော
ကိုကင်း	ထောင်းခနဲ	လေးလေးနက်နက်
ကိုက်တံ	လက်ခနဲ	ကျယ်ကျယ်ပြန့်ပြန့်

With this technique, morphological structure of words also can be analyzed. See in Table 3. The above-mentioned three and four-syllable words are adverbs derived from the verbs "ထူးဆန်း", "ထူးကဲ", and "ထူးခြား".

Statistical dictionaries can be updated much more easily than published printed dictionaries, which need more time, cost and man power to bring out a fresh edition. Common names such as names of persons, cities, committees etc. can be also mined.

Table 2. Words and Syllable Structure

No. of Syllables	No of words	Examples
1	2095	ကောင်း Good (Adj)
2	49825	လိပ်ပြာ Butterfly, Soul (N)
3	12147	ပြတင်းပေါက် Window (N)
4	10541	ပြည်တွင်းထုတ်ကုန် Domestic Product (N)
5	6242	လျှပ်စစ်ထမင်းအိုး Rice Cooker(N)
6	3314	သူနာပြုဆရာမ Nurse(female) (N)
7	1545	ရူပဗေဒပညာရှင် Physicist (N)
8	726	ပြည်ထောင်စုမြန်မာနိုင်ငံတော် Union_of_Myanmar (N)
9	301	သဲယံဧာတအရင်းအမြစ်များ Natural_Resources (N)
10	1137	ဧရာဝတီမြစ်ဝကျွန်းပေါ်ဒေသ Irrawady_ Delta_Region (N)

5. Conclusion and Further Work

Since words are not uniformly delimited by spaces in Myanmar script, segmenting sentences into words is an important task in NLP. In this paper we have described the need and possible techniques for segmentation in Myanmar script. In particular, we have used a combination of stop word removal and syllable level n-grams to hypothesize valid words with about 65% accuracy. Necessary scripts have been written in Perl. Manual checking is then performed to build lists of valid words. We have collected about 100,000 words for Myanmar including closed class words. We have used fairly simple and intuitive methods to build our initial data but now we can in fact use more powerful statistical and machine learning techniques to improve the performance of word hypothesizer and build larger and more comprehensive dictionaries and other lexical resources. We have mentioned the advantages of statistical methods and highlighted further work that can now be taken up. We hope this work will help to accelerate work in Myanmar language and larger lexical resources will be developed soon.

Table 3. Morphological Analysis of Myanmar

A 1-syllabe word	B(verb)= A+ "သည်"	C(noun)= "အ"+A	D(negative)= "မ"+A+ "ဘူး"	E (noun)= A+ "မှု"
ကောင်း good	ကောင်းသည် good	အကောင်း good	မကောင်းဘူး not good	ကောင်းမှု good deeds
ဆိုး bad	ဆိုးသည် bad	အဆိုး bad	မဆိုးဘူး not bad	ဆိုးမှု bad deeds
ရောင်း sell	ရောင်းသည် sell	အရောင်း sale	မရောင်းဘူး Not sell	ရောင်းမှု Sale
ရေး write	ရေးသည် write	အရေး writing	မရေးဘူး Not write	ရေးမှု writing
ပြော talk	ပြောသည် talk	အပြော talk	မပြောဘူး Not talk	ပြောမှု talk,talking

Table 4. English Stop Words Vs Myanmar Stops Words

English	Myanmar
Prepositions and adverbs	
above, among always, between, before, beside, down, inside, never, often, quite, while, with	အပေါ်၊ အနက်၊ အမြဲတမ်း၊ အတွင်းတွင်၊မကြာမီ၊ မတိုင်မီ၊ဒါ့အပြင်၊ အောက်မှာ၊ အထဲမှာ၊ ဘယ်တော့မျှ၊ မကြာခဏ၊ တော်တော်လေး ၊စဉ်တွင်၊ နှင့်အတူ၊ နှင့်၊ နှင့်တကွ
Nominative personal pronouns	
I, you , he, she, it, we, you, they	ကျွန်တော်၊ ကျွန်မ၊ ငါ၊ ကျုပ်၊ ကျွန်ုပ်၊ ကျနော်၊ ကျမ၊ သူ၊ သူမ၊ ထိုဟာ၊ ထိုအရာ၊ ဤအရာ၊ ထို၊ ၎င်း၊ ကျွန်တော်တို့၊ ကျွန်မတို့၊ငါတို့၊ ကျုပ်တို့၊ ကျွန်ုပ်တို့၊ ကျနော်တို့၊ ကျမတို့၊ သင်၊ သင်တို့၊ နင်တို့၊ မင်း၊ မင်းတို့၊ သူတို့
Nominative personal pronouns	
I, you , he, she, it, we, you, they	ကျွန်တော်၊ ကျွန်မ၊ ငါ၊ ကျုပ်၊ ကျွန်ုပ်၊ ကျနော်၊ ကျမ၊ သူ၊ သူမ၊ ထိုဟာ၊ ထိုအရာ၊ ဤအရာ၊ ထို၊ ၎င်း၊ ကျွန်တော်တို့၊ ကျွန်မတို့၊ငါတို့၊ ကျုပ်တို့၊ ကျွန်ုပ်တို့၊ ကျနော်တို့၊ ကျမတို့၊ သင်၊ သင်တို့၊ နင်တို့၊ မင်း၊ မင်းတို့၊ သူတို့
Accusative personal pronouns	
me, you, him, her, it, us, you, them	ကျွန်တော်အား၊ ကျွန်တော်ကို၊ ကျွန်မကို၊ ငါကို၊ ကျုပ်ကို၊ ကျွန်ုပ်ကို၊ သူ့ကို၊ သူမကို၊ ထိုအရာကို၊ သင့်ကို၊ သင်တို့ကို၊ နင်တို့ကို၊ မင်းကို၊ မင်းတို့ကို၊ ငါတို့ကို၊ ကျုပ်တို့ကို၊ ကျွန်ုပ်တို့ကို
Reflexive personal pronouns	
myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself	မိမိကိုယ်တိုင်၊ မိမိဘာသာ၊ မင်းကိုယ်တိုင်၊ မင်းဘာသာ၊ မင်းတို့ကိုယ်တိုင်၊ မင်းတို့ဘာသာ၊ သူကိုယ်တိုင်၊ ကိုယ်တိုင်၊ သူမကိုယ်တိုင်၊ သူ့ဘာသာ၊ သူ့ကိုယ်ကို၊ ကိုယ့်ကိုယ့်ကို၊ မိမိကိုယ်ကို၊ ၎င်းပင်၊ ထိုအရာပင်
Relative pronouns	
That	သည့်၊ မည့်၊ တဲ့
Possessive pronouns and adjectives	
my , your, his, her, its, our, your, their, mine, yours, his, hers, ours, yours, theirs	ကျွန်ုပ်၏၊ ကျွန်တော်၏၊ ကျွန်မ၏၊ ကျနော်၏၊ ကျမ၏၊ သူ၏၊ သူမ၏၊ ထိုအရာ၏၊ ထိုဟာ၏၊ ကျွန်ုပ်တို့၏၊ ငါတို့၏၊ ကျွန်တော်တို့၏၊ ကျွန်မတို့၏၊ ကျနော်တို့၏၊ ကျမတို့၏၊ သင်၏၊ သင်တို့၏၊ မင်း၏၊ မင်းတို့၏၊ သူတို့၏၊ ကျွန်တော့်ဟာ၊ ကျွန်မဟာ၊ ကျနော်၏ဟာ၊ ကျမဟာ၊ သူ့ဟာ၊ သူမဟာ၊ ကျွန်ုပ်တို့ဟာ၊ ကျွန်တော် တို့ဟာ၊ ကျမတို့ဟာ၊ သင်တို့ဟာ၊ မင်းတို့ဟာ၊ သူတို့ဟာ
Demonstrative pronouns and adjectives	
this, that, these, those	ဤအရာ၊ ဟောဒါ၊ ဟောဒီ၊ ထိုအရာ၊ ၎င်းအရာ၊ ယင်းအရာ၊ အဲဒါ၊ ဟိုဟာ
Indefinite pronouns and adjectives	
some , any, no, none, other, another, every, all, others, each, whole, both, neither, someone, somebody, something, anyone, anybody, anything, nobody, nothing, everyone, everybody, everything	အချို့၊ တစ်ခုခု၊ အဘယ်မဆို၊ ဘယ်အရာမဆို၊ အဘယ်မည်သော၊ အကြင်၊ အရာရာတိုင်း၊ စိုးစဉ်မျှ၊ ဘယ်လောက်မဆို၊ တစုံတရာ၊ အလျဉ်းမဟုတ်၊ မည်သည့်နည်းနှင့်မျှမဟုတ်၊အလျဉ်းမရှိသော၊ အခြားဖြစ်သော၊ အခြားသော၊ အခြားတစ်ခု၊ အခြားတစ်ယောက်၊ အားလုံး၊ အရာရာတိုင်း၊ အကုန်လုံး၊ အလုံးစုံ၊ အရာခပ်သိမ်း၊ တစ်ခုစီ၊ အသီးသီး၊ တစ်ဦးဦး၊ တစ်ခုခု၊ ကိုယ်စီကိုယ်ငှ၊ကိုယ်စီ၊ တစ်ဦးစီ၊ တယောက်စီ၊ တစ်ခုစီ၊ အကုန်၊ အပြည့်အစုံ၊ လုံးလုံး၊ နှစ်ခုလုံး၊ နှစ်ယောက်လုံး၊ နှစ်ဘက်လုံး၊ တစ်စုံတစ်ရာ၊ တစုံတခု၊ တစ်စုံတစ်ခု၊ တစ်စုံတစ်ယောက်၊ တစ်ယောက်ယောက်၊ မည်သူမဆို၊ ဘာမျှမရှိ၊ လူတိုင်း၊ လူတကာ၊ အဘယ်အရာမျှမရှိ

Conjunctions	
and, or, but, because, if, as, such	နှင့်၊ ပြီးလျှင်၊ ၎င်းနောက်၊ သို့မဟုတ်၊ သို့တည်းမဟုတ်၊ သို့မဟုတ်လျှင်၊ ဒါမှမဟုတ်၊ ဖြစ်စေ၊ သို့စေကာမူ၊ ဒါပေမယ့်၊ ဒါပေမဲ့၊ မှတစ်ပါး၊ မှလွဲလျှင်၊ အဘယ်ကြောင့်ဆိုသော်၊ သောကြောင့်၊ သဖြင့်၊ ၍၊ သည့်အတွက်ကြောင့်၊ လျှင်၊ ပါက၊ အကယ်၍၊ သော်ငြားလည်း၊ စေကာမူ၊ နည်းတူ၊ ပေမဲ့၊ ထိုနည်းတူစွာ၊ ကဲ့သို့၊ နှင့်စပ်လျဉ်း၍။ ယင်းကဲ့သို့၊ ထိုကဲ့သို့၊ ဤမျှ၊ အခုလောက်ထိ၊ ဤမျှလောက်၊ ဤကဲ့သို့၊ ဒါကတော့၊
Questions	
how, who, why, what, where, whose, when, whom, which	အဘယ်ကဲ့သို့၊ မည်ကဲ့သို့၊ မည်သည့်နည်းနှင့်၊ မည်သည့်နည်းဖြင့်၊ မည်သို့၊ ဘယ်လိုလဲ၊ သို့ပေမည့်၊ မည်သည့်နည်းနှင့်မဆို၊ ဘယ်နည်းနှင့်၊ မည်ရွေ့မည်မျှ၊ အဘယ်မျှလောက်၊ ဘယ်လောက်၊ မည်သူ၊ ဘယ်သူ၊ မည်သည့်အကြောင်းကြောင့်၊ ဘာအတွက်နဲ့လဲ၊ ဘာကြောင့်၊ မည်သည့်အတွက်ကြောင့်၊ အဘယ်ကြောင့်၊ မည်သည်၊ ဘာလဲ၊အဘယ်အရာနည်း၊ မည်သည့်အရပ်မှာ၊ဘယ်နေရာတွင်၊ မည်သည့်နေရာတွင်၊ မည်သည့်နေရာသို့၊ ဘယ်နေရာသို့၊ ဘယ်နေရာမှာ၊ ဘယ်သူ၏၊ မည်သည့်အရာ၏၊ မည်သည့်အခါ၊ ဘယ်အချိန်၊ ဘယ်အခါ၊ မည်သည့်အချိန်၊ ဘယ်တော့၊ မည်သူကို၊ ဘယ်သူ့ကို၊ မည်သူမည်ဝါ၊ မည်သည့်အရာ
Other (pronouns, prepositions)	
however, whoever, whatever, wherever, whenever, whomever	မည်သို့ပင်ဖြစ်စေ၊ မည်ရွေ့မည်မျှဖြစ်စေ၊ မည်သည့်နည်းနှင့်မဆို၊ ဘယ်လိုပဲဖြစ်ဖြစ်၊ မည်သူမဆို၊ ဘယ်သူမဆို၊ အဘယ်သူမဆို၊ မည်သည့်အရာမဆို၊ ဘာဖြစ်ဖြစ်၊ မည်သည့်အရာဖြစ်ဖြစ်၊ မည်သည့်အရပ်၌မဆို၊ မည်သည့်နေရာမဆို၊ ဘယ်အခါမဆို၊ ဘယ်အချိန်မဆို၊ ဘယ်အခါဖြစ်ဖြစ်၊ အချိန်အခါမရွေး၊

Table 5. Segmentation: Input and output

1	ပန်းခြံထဲတွင်လူတစ်ယောက်ရှိနေသည် ပန်းခြံ ထဲတွင် လူ တစ်ယောက် ရှိနေသည်
2	ဇီးကွက်အားလုံးသည်အမှောင်ထဲတွင်မြင်နိုင်သည် ဇီးကွက် အားလုံး သည် အမှောင် ထဲတွင် မြင်နိုင်သည်
3	ချိန်းဆိုချက်မတိုင်မီသူရောက်လာသည် ချိန်းဆိုချက် မတိုင်မီ သူ ရောက်လာသည်
4	ကျောင်းနေစဉ်ကကျွန်တော့်ဝါသနာမှာပြေးခုန်ပစ်အားကစားဖြစ်သည် ကျောင်းနေ စဉ်က ကျွန်တော့် ဝါသနာ မှာ ပြေးခုန်ပစ် အားကစား ဖြစ်သည်
5	သူဘောင်းဘီရှည်ပွပွဝတ်တတ်သည် သူ ဘောင်းဘီရှည် ပွပွ ဝတ်တတ်သည်
6	ကောင်းကင်တွင်သက်တန့်သည်လေးကိုင်းသဏ္ဌာန်ဖြစ်နေသည် ကောင်းကင် တွင် သက်တန့် သည် လေးကိုင်း သဏ္ဌာန် ဖြစ်နေသည်
7	ကစားနည်းအလိုက်သတ်မှတ်ထားသောစည်းကမ်းများကိုလိုက်နာကြရမည် ကစားနည်း အလိုက် သတ်မှတ်ထား သော စည်းကမ်းများ ကို လိုက်နာ ကြရမည်
8	မီးသတ်တပ်ဖွဲ့ရောက်လာချိန်တွင်အဆောက်အဦးမှာမီးအကြီးအကျယ်တောက် လောင်လျက်ရှိသည် မီးသတ်တပ်ဖွဲ့ ရောက်လာချိန်တွင် အဆောက်အဦး မှာ မီး အကြီးအကျယ် တောက်လောင် လျက် ရှိသည်
9	သူ၏ဝတ်စားဆင်ယင်ပုံမှာအခါကာလ၊နေရာဒေသနှင့်လျော်ကန်သည် သူ၏ ဝတ်စားဆင်ယင်ပုံ မှာ အခါ ကာလ ၊ နေရာဒေသ နှင့် လျော်ကန်သည်
10	သူမရှိကြောင်းကိုသိသွားသည် သူမ ရှိ ကြောင်းကို သိသွားသည်
Observe that we can obtain new words (Eg. “ဝတ်စားဆင်ယင်ပုံ”see item 9). “အခါ ကာလ” is over segmentation and showing that it should be included in dictionary. In item 10, because of longest word matching, the program incorrectly segments (he,“သူ”) to (she,“သူမ”). The negative sentence turns to positive.	

References

[1] Thomas Emerson**,** "Segmenting Chinese in Unicode", *16th International Unicode conference*, Amsterdam, The Netherlands, March 2000.

[2] Kavi Narayana Murthy, "Natural Language Processing - an Information Access Perspective", In Print.

[3] Myanmar Grammar Text Books 5th – 10th standards, Ministry of Education, Union of Myanmar.

[4] Saya U Soe, "The Khit Thit English-English- Myanmar Dictionary with Pronunciation", Yangon, Myanmar, April 2000.

[5] Myanmar-English Dictionary, Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

[6] Paulette Hopple, "The structure of nominalization in Burmese", *Ph.D Thesis*, May 2003.

[7] English-English-Myanmar Dictionary, Ministry of Commerce, CD version, Version 1.

[8] English- Myanmar Dictionary, Ministry of Education, CD version.

[9] Angell, R., Freund, G., and Willet, P. "Automatic spelling correction system using a trigram similarity measure ". Information Processing & Management, Information Research, Vol. 19 No. 4, January 2001.

[10] Pirkola, A, Keskustalo, Heikki, Leppänen, Erkka, Känsälä, Antti-Pekka and Järvelin, Kalervo, "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants", Information Research, Vol. 7 No. 2, January 2002.

[11] http://www.mcf.org.mm/unicode/doc/20010907_myanmar _syllables.pdf

[12] Vlado Keselj, "Text ::Ngrams" software, http://search.cpan.org/~vlado/ Text-Ngrams-1.8/

[13] http://www.syger.com/jsc/docs/stopwords/english.htm