ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Automatic Syllabification Rules for Manipuri Language

Mayanglambam Premi Devi
Department of Computer Science, Manipur University
Canchipur, India

Irengbam Tilokchan Singh
Department of Computer Science, Manipur University
Canchipur, India

Dr. Haobam Mamata Devi
Department of Computer Science, Manipur University
Canchipur, India

Abstract: This paper presents a rule-based automatic syllabification for Manipuri Language. Syllabification is the process of separation or identification or extraction of syllables in a word or in a sentence. Most of the syllabification tasks are done manually. The syllabification rules differ from one language to another as different languages have different syllable structures. Syllabification is the backbone of tasks like text-to-speech (TTS) conversion system, speech synthesizer, speech recognition, transliteration system etc. In this paper, we proposed an efficient algorithm for automatic syllabification of Manipuri Language based on syllable rule structure. The algorithm is evaluated on Manipuri words obtained from different sources like text-books, newspapers etc. The algorithm's result achieved 99.8% accuracy as compared to manual syllabification.

Keywords: Manipuri Language, phoneme, syllable, syllabification, monophthongs, diphthongs, text-to-speech

I. INTRODUCTION

In Natural Language Processing (NLP), the automatic syllabification process is an essential pre-requisite for speech synthesis systems like text-to-speech, speech recognition, etc and transliteration system. However, the task is non-trivial and over the last decade, several techniques have been added.[1] And there are many public resources for some languages (e.g., English, Finnish, Italian, Spanish and Japanese), but the resources for Manipuri Languages are still limited.

The automatic syllabification has two approaches, Rule-based and data-driven.[2-7] The rule-based approaches are the theoretical position of syllable whereas the data-driven approaches are based on examples which have been already syllabified. In this paper, we proposed an algorithm for automatic syllabification which works on rule-based, with some new proposals, especially in the treatment of words with diphthongs.

II. MANIPURI LANGUAGE AND ITS PHONOLOGICAL SYSTEM

Manipuri (/mənipuri/) is a language mainly spoken in the state of Manipur which is located at the North-Eastern region of India. And it is also spoken in the Indian states (Arunachal Pradesh, Assam, Meghalaya, Mizoram, Nagaland and Tripura) and in the countries like Myanmar and Bangladesh. It belongs to Tibeto-Burman language. The phonological system of Manipuri consists of three major systems of sound (vowels, consonants and tones) and one minor system of accompanying elements called juncture. The tones are of two type viz level and falling.[8-11]

There are 6 vowels in Manipuri. The vowels /ə/ and /a/, are distinguished, and are written as /a/ and /aa/. Table I shows the Manipuri vowels of 6 monophthogs {**/i/, /u/, /e/, /o/, /a/, /ə/**}. [8,12]

Table I: Vowels in Manipuri

	FRONT	***CENTRAL***	***BACK***
HIGH	**i**		**u**
MID	**e**	**ə**	**o**
LOW		**a**	

The vowel sounds in Manipuri Language occur in all the three positions (initially, medially and finally). Examples are shown in table II.

Table II: Examples of Vowel sound

	Initially	***Medially***	***Finally***
/i/	ipa 'father', ima 'mother', isiŋ 'water'	lin 'snake'	Hi 'boat', li 'cane'
/e/	ensaŋ 'curry'	len 'hail stone', ceŋ 'rice uncooked'	ce 'paper'
/ə/	əŋaŋ 'baby'	kəppa 'cry', səm 'hair', ləŋ 'thread'	əmə 'one', honbə 'to roar'
/a/	amuba 'black'	lan 'war', paw 'news'	ta 'spaer', ka 'room'
/o/	obə 'to vomit'	poŋ 'raft', koy 'beard'	thəro 'lily'
/u/	uci 'rat'	yum 'house'	yu 'alcohol'

Table III: Consonants in Manipuri

			Labial	***Coronal***	***Dorsal***	***Glottal***
Nasal			/m/	/n/	/ŋ/	
stops	***voiced***	***unaspirated***	/b/	/d/	/g/	
		breathy-voiced	/bh/	/dh/	/gh/	
	Voiceless	***unaspirated***	/p/	/t/	/k/	
		aspirated	/ph/	/th/	/kh/	
Fricative				/s/		/h/
Trill				/r/		
Lateral/Flap				/l/		
Approximant			/w/		/j/	

There are two semi-vowels in Manipuri viz /y/ and /w/. When these semi-vowels occur after the vowel, a diphthong like sound is produced. Diphthongs /əj/ and /əw/ are written as <ei> and <ou>. The commonly exhibit vowel's diphthongs are /ui/, /ao/, /ou/, /oi/, /ai/, /ei/, /əi/, /au/ and /əu/.

There are twenty-four consonant sounds including seven borrowed phonemes in Manipuri Language. We considered /ch/, /dh/, /gh/, /bh/, /jh/, /kh/, /ph/, /sh/, /th/ and /ng/ as single character. The consonants in Manipuri are shown in table III.[8]

Thus, the sound systems of Manipuri consist of twenty-four consonants and six vowels under segmental features.

III. SYLLABIFICATION

Syllabification is the process of separation or identification or extraction of syllables in a word from a sentence. If the word has only one syllable (i.e monosyllable) then, there is no need for syllabification. Thus, the process of syllabification is applicable only for the polysyllabic words. Polysyllabic word means word having more than one syllables i.e words having many phonological sounds. The words having two syllables are known as bi-syllable or di-syllable words. And tri-syllable word means word having three syllables. In this section, we will discuss the syllabification process of Manipuri language.

A. Syllable Structure of Manipuri

Manipuri language words have the tendency to reduce disyllabic form to monosyllabic nature. The monosyllabic or polysyllabic Manipuri words are co-occurred with either rising tones or falling tones.

The syllables are the sequences of phonemes in segments of the consonants(C) and the vowels (V) and also of the clusters of consonants and/or vowels. The syllable structure of Manipuri phonemes of vowels (V) and consonants (C) are as follows:-

i). V	/i/ 'blood', /i/'roof-cover-leaf'
ii). VC	/in/ 'fishing net',/un/ 'ice',
iii). CV	/hi/ 'boat', /ya/ 'teeth', /phi/ 'cloth'
iv). CVC	/khut/ 'hand', /kok/ 'head',
v). CCV	/kwa/ 'betal nut'
vi). CCVC	/kwak/ 'crow', /khwang/ 'waist'

Another syllable having the pattern CVCC is also found using in the words like Churachandpur /chand/, Səgolbənd /bənd/, /bəndh/ 'strike' etc.

B. Syllabification Rules

The rules for syllabification strictly follow as (i.) Every syllable has one vowel sound and (ii.) The number of vowel sounds equals the number of syllables. However a diphthong has two vowels. The /OW/ sound as in the word 'out' has two vowel sounds /ah/ and /oo/ that come together. Thus the above rules are not feasible in such syllables which have diphthongs. So, we treat the diphthong as single vowel unit. We used hyphen symbol to represent the syllable boundary in a word. The syllabification rules developed after studying disyllabic or bi-syllabic Manipuri words are given below:

(Observed sequence: Segmentation rule)

1) VCV:V-CV e.g ima(mother) : i-ma
2) VCVC:V-CVC e.g. ahəl(old person) : a-həl
3) VCVCC:V-CVCC e.g. Aizawl : Ai-zawl
4) VCCV:VC-CV e.g. unsa(skin) : un-sa
5) VCCVC:VC-CVC e.g.Imphal: Im-phal
6) CVVC:CV-VC e.g. saun(leather):sa-un
7) CVCV:CV-CV e.g. nabə (sick):na-bə
8) CVCVC:CV-CVC e.g. mə-təm(time):mə-təm
9) CVCVCC:CV-CVCC e.g. Thailand:Thai-land/
10) CVCCV:CVC-CV e.g. khutsa(fingers) :khut-sa
11) CVCCVC:CVC-CVC e.g.səm-chet(comb):səm-chet
12) CVCCCVC:CVCC-CVC e.g. Sanskrit:Sans-krit

The list of rules for tri-syllabic Manipuri words are given below:

(Observed sequence: Segmentation rule)

1) VCVCV:V-CV-CV
e.g. ətiya (sky):ə-ti-ya
2) VCVCVC:V-CV-CVC
e.g.isamak(myself):i-sa-mək
3) VCVCCV:V-CVC-CV
e.g. asəngba (green): a-səng-ba
4) VCVCCVC:V-CVC-CVC
e.g. Ibenpok(grandmother): i-ben-pok
5) VCCVCV:VC-CV-CV
e.g. okpəgi(reception):ok-pə-gi
6) VCCVCVC:VC-CV-CVC
e.g.unnapham (meeting point): un-na-pham
7) CVVCV:CV-V-CV
e.g. mioiba (human being): mi-oi-ba
8) CVCVCV:CV-CV-CV
e.g. samətu (wool):sa-mə-tu
9) CVCVCVC:CV-CV-CVC
e.g .mə-sa-mək (himself/herself): mə-sa-mək
10) CVCVCCV:CV-CVC-CV
e.g. miyamgi (of the people):mi-yam-gi
11) CVCVCCVC:CV-CVC-CVC
e.g. wa-hən-thok (meaning):wa-hən-thok
12) CVCVCCVCC:CV-CVC-CVCC
e.g. Səgolbənd (name of a place):sə-gol-bənd
13) CVCCVCV:CVC-CV-CV
e.g. yamləba (huge):yam-lə-ba
14) CVCCVCVC:CVC-CV-CVC
e.g. kunnipal(twenty eight) :kun-ni-pal
15) CVCCVCVCC:CVC-CV-CVCC
e.g. thangmeibənd (name of a place):thang-mei-bənd
16) CVCCVCCV:CVC-CVC-CV
e.g. ləmləkki(wild):ləm-lək-ki
17) CVCCVCCVC:CVC-CVC-CVC
e.g. pənthungphaəm (destination):pən-thung-phəm
18) CCVCVCVC:CCV-CV-CVC
e.g. kwakeithel(name of a place):kwa-kei-thel

19) CCVCVCCVCC:CCV-CVC-CVCC
    e.g. khwairəmbənd (name of a market): khwai-rəm-bənd
20) CCVCCVCV:CCVC-CV-CV
    e.g. kwaksiphai(name of a place): kwak-si-phai

C. Implementation

The above Segmentation rules were implemented in the syllabification algorithm. First the algorithm extracts all the distinct words from the input text. Then the segmentation rule of syllabification was applied to the distinct words. Before implementing the syllabification, first we find the C-V (Consonant-Vowel) cluster pattern of the distinct or given words. Then the corresponding C-V pattern was lookup in the mapping (hashing or dictionary) table with 'key' : 'value' as the 'Observed_sequence' : 'Segmentation_rule' of syllabification. If the C-V observed sequence is found as a key then the corresponding value is printed out as the resulting segmented syllables.

The procedure of the syllabification algorithm is shown as follow:

```
ALGORITHM:SYLLABIFICATION
SELECT the distinct words from the input text
WHILE(word in distinct words):
    DETECT the C-V pattern of the word
    IF(C-V pattern is in the key list of syllable hash-table)
     {      -then PRINT  the key-value mapping}
    ELSE
     {      -PRINT pattern not found.}
END WHILE
```

IV. Results And Discussions

The above syllabification algorithm was tested on 1, 24,087 distinct words which were extracted from a corpus of size 3000 pages. The result is compared with manual syllabification to measure the accuracy.

The algorithm achieves 99.8% accuracy with compared to manual syllabification by an expert. Since it implements hash table, the average and amortized case complexity is O(1) and the worse case complexity suffer O(*n*).

V. Conclusion

Automatic syllabification is an important but difficult problem that has implications on pronunciation generation for text-to-speech synthesis and transliteration system. There are essentially two possible approaches to automatic syllabification: rule-based and data-driven.

In this work, we have concluded that the rule-based and data-driven example based have thin margin to separate as the rules are derived from the examples of the previous knowledge. More variant data derive more rules.

VI. Limitation

The Manipuri words like Pangei (name of a place), kangi (houjik kangi 'nowadays'), sa-səngum (like wild animal), ingəni (natung ingəni 'will follow'), etc cannot be segmented because we considered /ch/, /dh/, /gh/, /bh/, /jh/, /kh/, /ph/, /sh/, /th/ and /ng/ as single character.

VII. References

[1] Nelson Neto, Willian Rocha, Gleidson Sousa, “An open-source rule-based syllabification tool for Brazilian Portuguese,” Journal of the Brazilian Computer Society 21(1): 1:1-1:10 (2015).

[2] Kimmo Kettunen, Paul McNamee,Feza Baskaya, “Using Syllables As Indexing Terms in Full-Text Information Retrieval,” Baltic(HLT), Frontiers in Artificial Intelligence and Applications, IOS Press, 2010, vol 219, pp. 225-232.

[3] C. R. Adsett, Y. Marchand, “A Comparison of Data-Driven Automatic Syllabification Methods,” SPIRE 2009, LNCS 5721, Saariselka, Finland, August 2009, pp. 174–181.

[4] C. R. Adsett, Y. Marchand and V. Keselj, “Syllabification rules versus data-driven methods in a language with low syllabic complexity: The case of Italian,” Computer Speech & Language, 23(4):444-463, 2009.

[5] Y. Marchand, C. Adsett and R.I. Damper, “Automatic Syllabification in English: A Comparison of Different Algorithms,” Language and Speech, 52(1):1-27, 2009.

[6] C.R. Adsett and Y. Marchand,” Are Rule-based Syllabification Methods Adequate for Languages with Low Syllabic Complexity? The Case of Italian,” SSW6-2007, Bonn, Germany, August 2007, pp.58-63.

[7] Y. Marchand, C.R. Adsett and R.I. Damper, “Evaluating Automatic Syllabification Algorithms for English,” SSW6-2007, Bonn, Germany, August 2007, pp.316-321.

[8] Chungkham Yashwanta Singh, “Manipuri Grammar” Rajesh Publication, 2000 ISBN: 81-85891-33-8.

[9] D.N.S. Bhat & M.S. Ningomba, “Manipuri grammar,” Published: Mu̇nchen: LINCOM Europa, 1997, ISBN: 3-89586-191-x.

[10] Shobhana Lakshmi Chelliah, “A Grammar of Meithei,” Published: New York : Mouton de Gruyter, 1997, ISBN: 3-11-014321-6, 3110143216, alk. Paper.

[11] Dr. Soibam Rebika Dev, “Is Manipuri an Endangered Language?”, Language in India. ISSN 1930-2940 Vol. 13:5 May 2013, pp520-433.

[12] Leihaorambam Sarbajit Singh, Soibam Imoba Singh, “Phonological Problems in Making English-Manipuri Dictionary for Manipuri Speakers,” Language in India vol 7 : 9 September 2007.