Automatic Syllabification for Manipuri language

Loitongbam Gyanendro Singh, Lenin Laitonjam, Sanasam Ranbir Singh
Computer Science and Engineering Department
Indian Institute of Technology Guwahati, Assam, India
Email: {gyanendrol9,lenin.lai,ranbir}@iitg.ernet.in

Abstract

Development of hand crafted rule for syllabifying words of a language is an expensive task. This paper proposes several data-driven methods for automatic syllabification of words written in Manipuri language. Manipuri is one of the scheduled Indian languages. First, we propose a language-independent *rule-based* approach formulated using *entropy* based phonotactic segmentation. Second, we project the syllabification problem as a sequence labeling problem and investigate its effect using various sequence labeling approaches. Third, we combine the effect of sequence labeling and rule-based method and investigate the performance of the hybrid approach. From various experimental observations, it is evident that the proposed methods outperform the baseline rule-based method. The entropy based phonotactic segmentation provides a word accuracy of 96%, CRF (sequence labeling approach) provides 97% and hybrid approach provides 98% word accuracy.

1 Introduction

Manipuri language, one of the scheduled Indian languages, belonging to a Tibeto-Burman languages family(Chelliah, 1990) is syllabic in nature. In general, a syllable follows *onset-nucleus-coda* (consonant-vowel-consonant) structure, where nucleus is the core component defined by a vowel. The preceding and following consonants defined by the onset and coda respectively may or may not be present in a syllable. However every syllable must have a nucleus. Formation of an onset, nucleus and coda, while producing an uninterrupted syllabic sound, greatly affects the pronunciation of a language. For various applications such as Text-to-Speech Synthesis (TTS) (Kishore and Black, 2003; Bellur et al., 2011), Automatic Speech Recognition (ASR) (Wu et al., 1998) etc., proper syllabification of a word is one of the important core issues, especially for syllabic Indian languages. Though there have been several studies in this direction for the rich resource Indian languages such as Hindi, Bengali, Tamil, Telegu, Malayalam, Marathi (Bellur et al., 2011; Narendra et al., 2011; Kurian et al., 2011), only very few studies have been reported for low resource Manipuri language (Abbi and Awadhesh, 1985; Chelliah, 1990). To the best of our knowledge, no corpus is available for Manipuri language in UTF-8 to study various aspects of corpus-based linguistic understanding. This has motivated us to generate Manipuri corpus in UTF-8 and analyse various characteristics such as onset and coda distributions, syllable formulation and syllabification rules etc.

Manipuri language is written generally using two scripts; *Bengali scripts* and *Meitei Mayek* [1]. At present, majority of the Manipuri documents are written using Bengali scripts and, hence the investigation in this paper focuses on Manipuri words written using Bengali script. Further, processing of Bengali scripts poses more challenges than processing Meitei Mayek because of the phoneme imbalance. Bengali script has 55 symbols to represent 38 phonemes in Manipuri language (Singh et al., 2007).

One of the major challenges in syllabifying Indian languages such as Hindi, Urdu, Bengali, Marathi, Kashmiri, Punjabi, Gujarati is the *schwa* deletion in their written forms (Kishore and Black, 2003;

[1] http://www.unicode.org/L2/L2000/00259-MeeteiMayek.pdf

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 349–357, Osaka, Japan, December 11-17 2016.

Choudhury et al., 2004). Majority of the existing studies for syllabifying Indian languages are based on hand crafted rules built on language characteristics. Building such rules requires domain knowledge and it is an expensive task. Further, such rules often fail to capture *heteronym* (word having same spelling, but different pronunciations).

In general, languages evolves over time. New vocabularies are added. One language borrows vocabularies from other languages and become parts of the regular usage. For example, words like `institution`, `election`, `daddy` etc. have become more preferred words than their Manipuri counterparts. While syllabifying such borrowed or loan words, the rules crafted using the characteristics of the native language often fail to capture the syllabic structures of such word.

In this paper, we first explored two rule-based approaches which are *Baseline System* (C*VC* structured) and *Entropy Based Phonotactic Segmentation* (EPS). Secondly, we transform the syllabification problem as a *sequence labeling problem* and investigate the effect of two state-of-art methods, namely *Conditional Random Field* (CRF) (Sutton and McCallum, 2006) and *Maximum Entropy Markov Model* (MEMM) (McCallum et al., 2000). The intuition behind such transformation is that with an appropriate training set, these models can learn the inherent dependencies automatically and address the problem of *schwa*, *heteronym* and *loan* words. Another advantage of such machine learning methods is the language independent model. Given an appropriate annotated dataset, these methods can syllabify with reasonable accuracy without the knowledge of the underlying language. In summary, this paper has the following five contributions.

- Generate Manipuri corpus in UTF-8.
- Statistical analysis of the syllabified corpus for understanding syllable components (i.e. onset, nucleus and coda).
- Development of a data driven rule-based syllabification methods by exploring entropy distribution.
- Investigating the syllabification performance of CRF, MEMM and comparing with rule-based counterparts.
- Evaluating CRF model over Assamese language dataset.

2 Related Works

Two approaches are mainly considered for automatic syllabification of a language.

(i) **Rule-based approach**

(a) **Obligatory Onset Principle** (Hooper, 1972), or **Principle of Maximum Open Syllabicity** (Pulgram, 1970): It is one of the simplest principles, which is based on the assumption that open syllable i.e., a syllable with no coda, is significance. This is in fact very naive approach and it is impractical for Manipuri as it has various legal consonant clusters other that just CV syllable structure.

(b) **Sonority principles** (Selkirk, 1984): In this method, syllabification depends on sonority of each phoneme i.e., its quality of being sonorous. Sonorous is the capability of giving out sound especially resonance, deep sound. This principle assigns numerical values to every phoneme of a syllable depending on its sonority level where vowels have the highest value followed by nasals, fricatives, and plosives. The main disadvantage of sonority information is the position dependency of phonemes. The same phoneme present at a different location of a syllable may have different sonorous property. There exists several instances where phoneme does not fit its typical sonority rules.

(c) **The legality principle** (Goslin and Frauenfelder, 2001): Syllabification is based on the validation of the syllable structure considering its onsets and codas. It also considers the legality of the consonant clusters to detect its splitting points. It allows consonant clusters to be valid if they appear in some syllables. Legality principle requires a big corpus to study the legality of its constituents structures. Its main problem arises when several valid splitting instances of a consonant cluster are possible resulting on ambiguous syllabification rules.

(d) **Maximum onset principle** (MOP) (Kahn, 2015): It is very similar to legality principle. Here, if multiple legal splits are possible, it gives preference to longer onsets irrelevant of the legality of the syllable coda.

Rule-based approaches are language dependent. A prior knowledge of the syllable structure and phonotactics of the language is necessary to derive such rules. Further, syllabification purely based on rule-based approached is mostly inadequate due to the presence of various ambiguities. There are various instances where correct syllabification cannot be obtained by a definite rule or may even break the conventional syllabification principle.

(ii) **Data driven approach**

(a) Zhang and Hamilton (1997) suggested Learning English Syllabification Rules system that learn rules using a symbolic pattern recognition approach.

(b) Adsett and Marchand (2009) provide a comparison between various data-driven syllabification algorithms namely, IB1 (94.36%) and Look-up Procedure (91.44%) algorithms along with Hidden Markov Support Vector Machines (95.17%), Liangs algorithm (95.48%) and Syllabification by Analogy (96.70%) which incorporate structure information.

(c) (Marchand et al., 2009; Adsett and Marchand, 2009) showed that rule-based approaches perform poorly as compared to the data driven approaches.

To the best of our knowledge, there have not been any work related to corpus-based machine learning approach using sequence labeling method for syllabifying Manipuri words. In the similar line of approach, Dinu et al. (2013) used CRF using two level tagging (phone boundary and phone distance) to perform syllabification for Romanian language. However, unlike their study, our approach uses only one level tagging as discussed in the section 5.2.

Further, Rogova et al. (2013) proposed a language-independent probabilistic syllabification of phonetic transcriptions of words using Segmental Conditional Random Fields (SCRF). This method works with phonetic transcriptions of word as input. Due to the existance of inherent $schwa$ in the script, phonetic transcriptions of a word may not be correct for Manipuri words.

3 Corpus generation

One of the challenges working with low resource Manipuri language is the unavailability of UTF-8 data sources. To create UTF-8 textual corpus for this study, we have identified two Manipuri local newspapers [2]. The local Manipuri newspapers written in Bengali script are not unicode compatible. The news articles are published and archived on the news website in PDF format. In order to generate unicode compatible text corpus, we deploy a Bengali OCR [3]. However, the accuracy of the OCR is very poor i.e., 52.6% words accuracy. We, therefore, manually correct the text extracted from the OCR. Thus, our corpus consists of 733 documents with 89,084 words and 26,203 unique words covering all possible Manipuri phone types.

4 Statistical analysis of Manipuri syllables

The 26,203 unique words present in the corpus are manually syllabified to study their characteristics. Each syllable is annotated with its positional information (`beg`, `mid` and `end`). For each word, the beginning syllable is marked as `beg`, ending syllable as `end` and all syllables that are between `beg` and `end` as `mid`. For example, the word *particular* is divided into *par*_`beg`, *ti*_`mid`, *cu*_`mid` and *lar*_`end`. Analysis of the distribution of onset, nucleus, and coda over different syllabic positions is important for understanding various linguistic aspects of a language. For the proposed rule-based approach for

[2] `http://www.poknapham.in/`, `http://naharolgithoudang.in/`

[3] `https://www.newocr.com/`

Table 1: Types of syllable structure with their positional distribution (in percentage)

Syllable structures	beg	mid	end	total
CV	12.61628	24.82558	21.98335	59.42521
CVC	12.13398	10.99894	2.97172	26.10465
CVV	1.72568	1.557875	0.56950	3.85306
CVCC	1.96749	1.346458	0.47040	3.78435
V	3.10253	0.202167	0.38583	3.69054
CCV	0.12552	0.447938	0.72013	1.29360
VC	0.71353	0.066067	0.04756	0.82716
CCVC	0.18498	0.121563	0.12684	0.43339
VV	0.31580	0.046247	0.00924	0.37130
CVVC	0.03699	0.005285	0.01453	0.05681
CCVV	0.02906	0.001321	0.00264	0.03303
CCVCC	0.02378	0.002642	0.00528	0.03171
VCC	0.02246	0.00	0.00	0.02246
CVCCC	0.00924	0.00	0.00528	0.01453
CVVCC	0.01189	0.00264	0.00	0.01453
CCVVC	0.00792	0.00	0.00132	0.00924
CCVVCC	0.00264	0.00264	0.00132	0.00660
CVVC	0.00264	0.00396	0.00	0.00660
CCCC	0.00528	0.00	0.00132	0.00660
VVC	0.00132	0.00132	0.00132	0.00396
VVCC	0.00264	0.00132	0.00	0.00396
CCVCCC	0.00264	0.00	0.00	0.00264
CCCV	0.00132	0.00	0.00132	0.00264
VCCC	0.00132	0.00	0.00	0.00132

syllabifying Manipuri word, we study distributional characteristics of onset, nucleus, and coda as one of the core parameters.

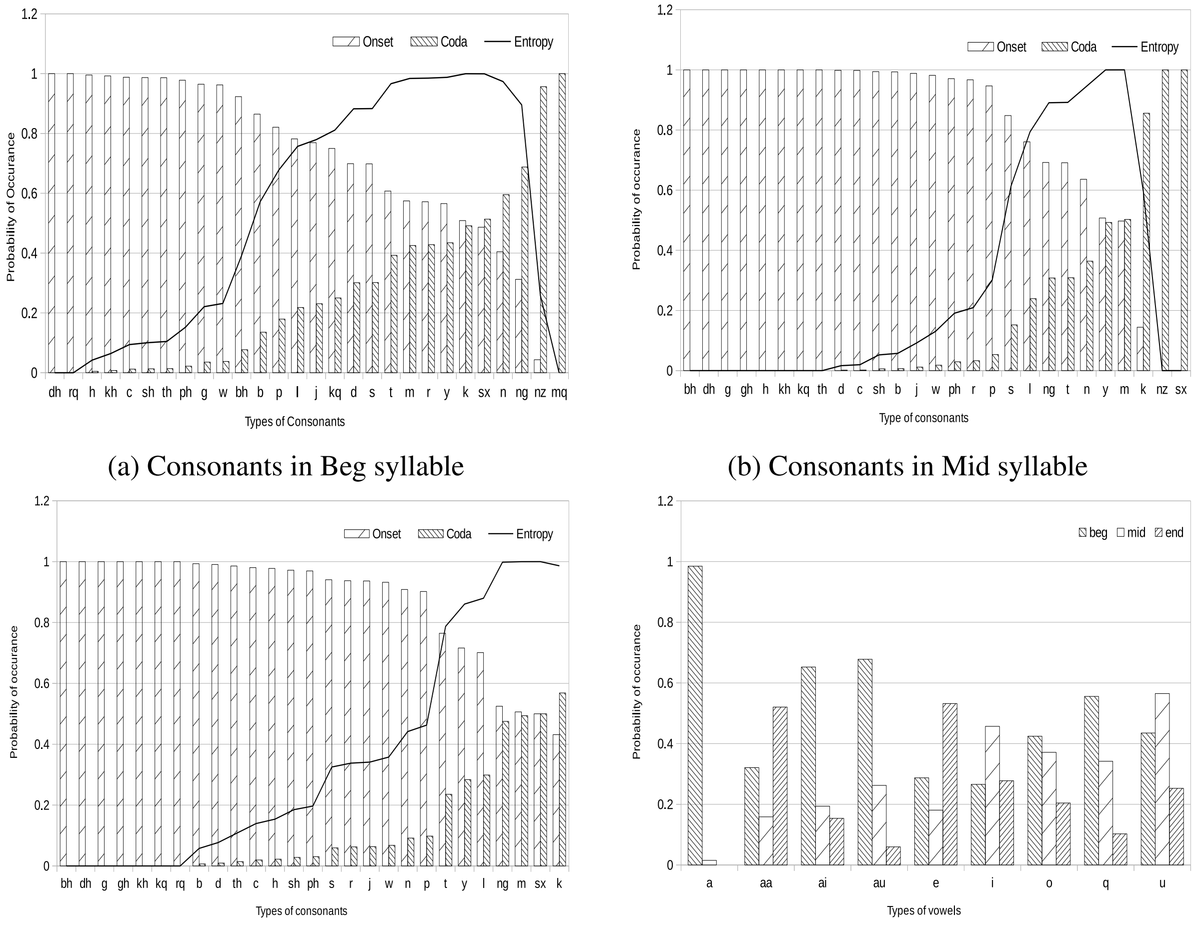
Table 1 shows the types of syllable structure with their percentage distribution of syllabic position (i.e. `beg`, `mid`, `end`) present across words in the corpus where C and V represent consonant and vowel respectively. The dashed line is the partition between the typical 9 syllable structures of Indian languages (covering almost 99.78%) over the nontypical syllable structures found in this corpus. All the root patterns of Manipuri language, as describe in Abbi and Awadhesh (1985), are found in the corpus. The appearance of the nontypical structures are due to the presence of loan words such as *headquarters*, *match*, *annual*, etc.

In Table 1, we observed that CV structure is the most commonly used syllabic structure with more than 59.43% of occurrences, followed by CVC with about 26.1%. We also observed that some syllable structures have dominant positional distribution. For example, V and CVC dominate with `beg`, while CV dominates with `mid` and CCV with `end` syllable position.

4.1 Phonotactics

Phonotactics is an important area of research in phonology which explores the rules governing the phoneme sequence of a language. To understand phonotactics of Manipuri language, we explore the distribution of onset, nucleus and coda across the syllables at different positions.

Figure 1[a,b,c] show the probability of onset and coda distribution for consonant phones at different syllabic positions (beg, mid and end) and their entropy i.e. the information contain in onset and coda probability distribution. A consonant with low entropy shows its structural bias toward either onset or coda. The figures show that a significant number of consonants have structural (onset or coda) as well as positional bias. Further, Figure 1[d] shows the probability distribution of nucleus vowel phones across different syllabic positions (beg, mid and end). It clearly shows that some vowels such as /a/, /au/, /ai/ have positional bias towards `beg`. In the proposed EPS method, discussed in section 5.1.2, the entropy



(a) Consonants in Beg syllable

(b) Consonants in Mid syllable

(c) Consonants in End syllable

(d) Vowels

Figure 1: Probability distribution of different phones

distribution is used to determine syllable boundary.

5 Proposed Methods

In this paper, we propose three different corpus-based automatic syllabification methods; (i) Rule-based approach, (ii) Sequence labelling approach and (iii) Hybrid of rule-based and sequence labelling approaches.

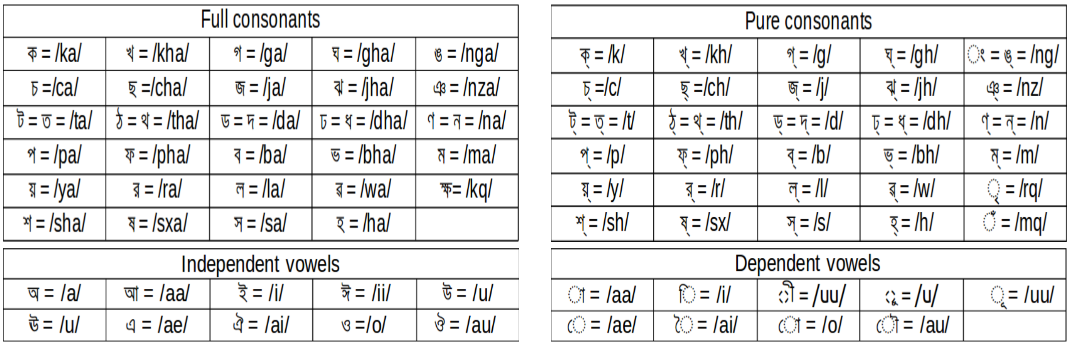
5.1 Rule-based approach

In this section, we discuss two rule-based approaches: (i) Purely based on C*VC* rule of the language and (ii) Corpus driven EPS. Both of them depends on the legality of syllable structures and consonant clusters by checking each phone along with its adjacent units until a legal split point is obtained. These two approaches differ only when a current phone α to be checked is consonant. When α is a dependent vowel and if its previous phone is a consonant then add α to the current syllable phone sequence else there is spelling error in the syllabifying word. When α is an independent vowel, then add it to the current syllable phone sequence only if it satisfies one of the following conditions.

a) If previous phone is one of the following short vowels {/a/, /aa/, /i/, /ii/} and followed by either /u/ or /uu/.

b) If previous phone is a short vowels /aa/ and followed by /o/.

c) If previous two phones do not satisfy condition (a) and previous phone is one of the following vowels {/a/, /aa/, /u/, /uu/, /o/} and followed by either /i/ or /ii/.



Full consonants				
ক = /ka/	খ = /kha/	গ = /ga/	ঘ = /gha/	ঙ = /nga/
চ =/ca/	ছ =/cha/	জ = /ja/	ঝ = /jha/	ঞ = /nza/
ট = ত = /ta/	ঠ = থ = /tha/	ড = দ = /da/	ঢ = ধ = /dha/	ণ = ন = /na/
প = /pa/	ফ = /pha/	ব = /ba/	ভ = /bha/	ম = /ma/
য় = /ya/	র = /ra/	ল = /la/	ৱ = /wa/	ক্ষ= /kq/
শ = /sha/	ষ = /sxa/	স = /sa/	হ = /ha/	

Pure consonants				
ক্ = /k/	খ্ = /kh/	গ্ = /g/	ঘ্ = /gh/	◌ং = ঙ্ = /ng/
চ্ =/c/	ছ্ =/ch/	জ্ = /j/	ঝ্ = /jh/	ঞ্ = /nz/
ট্ = ত্ = /t/	ঠ্ = থ্ = /th/	ড্ = দ্ = /d/	ঢ্ = ধ্ = /dh/	ণ্ = ন্ = /n/
প্ = /p/	ফ্ = /ph/	ব্ = /b/	ভ্ = /bh/	ম্ = /m/
য়্ = /y/	র্ = /r/	ল্ = /l/	ৱ্ = /w/	◌ৃ = /rq/
শ্ = /sh/	ষ্ = /sx/	স্ = /s/	হ্ = /h/	◌ঁ = /mq/

Independent vowels				
অ = /a/	আ = /aa/	ই = /i/	ঈ = /ii/	উ = /u/
ঊ = /u/	এ = /ae/	ঐ = /ai/	ও =/o/	ঔ = /au/

Dependent vowels				
◌া = /aa/	ি◌ = /i/	◌ী = /uu/	◌ু = /u/	◌ূ = /uu/
ে◌ = /ae/	ৈ◌ = /ai/	ে◌া = /o/	ে◌ৗ = /au/	

Table 2: Types of consonants and vowels and their orthographic phoneme representation

Otherwise, mark α as the beginning phone of the new syllable.

5.1.1 Baseline System

Since pure consonant cannot be a stand-alone phone, it must always be merged to either previous or following phone. Assumption in this approach is that a diacritic character (such as halant or virama) is properly placed to distinguish pure consonant from full consonant phone. Table 2 shows different types of consonants and vowels, and their corresponding orthographic representations. In this approach, if the input phone α is a pure consonant and not followed by one of the following semivowels $\{$/y/, /r/, /l/, /w/$\}$, then add α to the current syllable phone sequence, otherwise mark it as the beginning phone of the new syllable. We consider this approach as a *baseline* model to compare with other proposed methods.

5.1.2 Entropy Based Phonotactic Segmentation (EPS)

As the problem of schwa deletion exists in the script, the assumption made for the baseline system does not happen in reality. The proposed approach (described in algorithm 1), is an improvement over the baseline to handle the schwa deletion problem, by using the entropy estimate of phone and cluster of phones. It is designed as a generalized recursive algorithm where the baseline approach is treated as terminating case. Initially for each α, the two parameters β and γ, which are the preceding and following phones or string of phones of α, are taken as NULL value.

- H(x,y) calculate the cluster entropy of x and y when both are not NULL. If one of them is NULL, then the entropy of onset-coda distribution for the other is calculated.
- P_{split}(x,y) and P_{merge}(x,y) give the probability of splitting and merging respectively when x and y appear together.
- CheckSplit(x,y) function either split or merge x and y based on P_{split}(x,y) and P_{merge}(x,y) score.
- x^* represents the expansion of *x* by adding more context information.
- θ represents a threshold of the entropy.
- MAX = 5, since the typical syllable structures are not greater that 5.

5.2 Sequence labeling approach

In this section, we transform the problem of segmenting a word into its syllable units as a problem of sequence labeling task. To apply sequence labeling task on a word in the corpus, every letters present in each word are tagged as C or S where C denotes *continuous character* and *S* denotes *splitting character*. In each manually syllabified word, all letters except the last letter in a syllable are tagged as C and the last letter as S. For example, the word Manipuri which is syllabified as ma ni pu ri is tagged as m/C a/S n/C i/S p/C u/S r/C i/S.

```
EPS($\beta$,$\alpha$,$\gamma$)
if Length($\beta + \alpha + \gamma$) $\leq$ MAX then
    if H($\beta$,$\alpha$) $\leq \theta$ then
        CheckSplit($\beta$,$\alpha$)
    end
    else if H($\alpha$,$\gamma$) $\leq \theta$ and $P_{split}$($\alpha$,$\gamma$) $< P_{merge}$($\alpha$,$\gamma$) then
        EPS($\beta$,$\alpha$ + $\gamma$,NULL)
    end
    else
        EPS($\beta^*$,$\alpha$,$\gamma^*$)
    end
end
else
    Baseline($\alpha$)
end
```

Algorithm 1: EPS: If α is consonant

Now, considering the observed input sequence $o_1, o_2, ..., o_k$ (a Manipuri word), where o_i denotes an alphabet of Manipuri language (Bengali script in this study), the task is to find the best state sequence i.e., $s_1, s_2, ..., s_k$ that might have resulted the observed sequence where $s_i \in \{C, S\}$. For the word `Manipuri`, the desire output is the sequence $\{$`C S C S C S C S`$\}$.

In this study, we have considered two state-of-art sequence labeling methods namely CRF (Sutton and McCallum, 2006) and MEMM (McCallum et al., 2000). For this experiment, each phone that occurs less than 5 times in the corpus is considered as a rare phone. Each phone that occurs more than 2 times in the corpus will be a feature with their corresponding tags. Those features having a frequency less than 2 will be ignored. The rare phone features having a frequency less than 10 will be ignored and the common phone where feature frequency is more than 250 will form an equivalent class. For improving iterative scaling, 100 iterations have been used and to avoid the overfitting problem L1 regularizer have been employed.

5.3 Hybrid approach

The sequence labeling approach does not verify its output syllable structure with the phonotactic structure of the underlying language. In the proposed hybrid approach, the output of the sequence labeling method is passed to the proposed EPS method to verify the phonotactic structure and rectify the possible error.

6 Syllabification Results

This section discusses the experimental results of different syllabification methods proposed in this paper. 5-fold cross validation has been used to investigate the performance of different labeling methods. *Precision*, *Recall*, *F-measure* and *word accuracy* have been used as the evaluation metrices.

Table 3 compares the performance of different methods in terms of word accuracy, Precision, Recall and F1 measure and shows the distribution of wrongly syllabified words among the loan words and Manipuri words. It clearly shows that all the proposed methods outperform the baseline method. The EPS outperforms the baseline at least by 8.5% in accuracy. Similarly, CRF outperforms both baseline and EPS of about 9.6% and 1% respectively. Interestingly MEMM underperforms all other methods because of the existance of structural bias i.e. label bias problem. EPS and CRF can overcome this problem because they can generalize the global context information. It is also evident from the table that hybrid approach further enhanced the classification accuracy over CRF and EPS of about 0.5% and 1.6%. All the proposed approaches in this paper, were able to handle most of the schwa deletion and loan words problems.

Table 3: Evaluation for different methods

Manipuri Dataset				
Methods	Word Accuracy	Precision	Recall	F1 Score
Baseline	0.875	0.92950	0.93577	0.93212
EPS	0.95992 (+0.085)	0.98007	0.98004	0.98
MEMM	0.73225 (-0.143)	0.80968	0.80519	0.80603
CRF	0.97080 (+0.096)	0.98293	0.98343	0.98287
Hybrid	0.97576 (+0.101)	0.98624	0.98731	0.98653

Assamese Dataset				
Methods	Word Accuracy	Precision	Recall	F1 Score
MEMM	0.33808	0.55486	0.54891	0.54817
CRF	0.95258	0.97798	0.97777	0.97745

Evaluation for different types of word in Manipuri Dataset					
Methods	Word	Word			
	type	accuracy	Precision	Recall	F1 Score
Baseline	Loan	0.3227	0.6218	0.6518	0.6339
	Manipuri	0.8909	0.9389	0.9443	0.9412
EPS	Loan	0.7773	0.8791	0.9027	0.8890
	Manipuri	0.9657	0.9846	0.9825	0.9832
MEMM	Loan	0.2182	0.4241	0.4406	0.4289
	Manipuri	0.8100	0.8869	0.8814	0.8828
CRF	Loan	0.7864	0.8874	0.8950	0.8888
	Manipuri	0.9772	0.9868	0.9867	0.9865
Hybrid	Loan	0.8182	0.8961	0.9117	0.9022
	Manipuri	0.9805	0.9894	0.9897	0.9894

6.1 Language Independency

To investigate the capability of CRF handling across different languages, we have further considered another publicly available Assamese dataset consisting of 17,925 unique words. Second part of the Table 3 shows the response of CRF and MEMM over this dataset. It clearly shows that CRF provides a word accuracy of 0.95, while MEMM provides just 0.34.

7 Conclusion

In this paper, we first present generation of Manipuri text corpus using Bengali OCR for linguistics studies. Using the corpus generated from the news articles, we analysed phone characteristics across structural and positional properties. Using the observation from this study, we proposed entropy based phonotactic segmentation (EPS) method (a corpus driven rule based automatic syllabification method) for automatic segmentation of Manipuri words. This approach achieves a word accuracy of 96%. Further, the effect of CRF has been investigated and found to provide a word accuracy of 97%. Then, we combine both the CRF and EPS to enhance the classification accuracy and achieve an accuracy of 98%.

8 Acknowledgment

Authors would like to thank Miss Nanaobi Huidrom, Department of CSE, IIT Guwahati, for her support during corpus generation. This work is funded by Deity, Government of India.

References

Anvita Abbi and K. Mishra Awadhesh. 1985. Consonant clusters and syllabic structures of meitei. *Linguistics of the Tibeto-Burman Area*, 8(2):81–92.

Connie R Adsett and Yannick Marchand. 2009. A comparison of data-driven automatic syllabification methods. In *International Symposium on String Processing and Information Retrieval*, pages 174–181. Springer.

Ashwin Bellur, K Badri Narayan, K Raghava Krishnan, and Hema A Murthy. 2011. Prosody modeling for syllable-based concatenative speech synthesis of hindi and tamil. In *Communications (NCC), 2011 National Conference on*, pages 1–5. IEEE.

Shobhana L Chelliah. 1990. Level-ordered morphology and phonology in manipuri. *Linguistics of the Tibeto-Burman Area*, 13(2):27–72.

Monojit Choudhury, Anupam Basu, and Sudeshna Sarkar. 2004. A diachronic approach for schwa deletion in indo aryan languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, pages 20–26. Association for Computational Linguistics.

Liviu P Dinu, Vlad Niculae, and Octavia-Maria Sulea. 2013. Romanian syllabication using machine learning. In *Text, Speech, and Dialogue*, pages 450–456. Springer.

Jeremy Goslin and Ulrich H Frauenfelder. 2001. A comparison of theoretical and human syllabification. *Language and Speech*, 44(4):409–436.

Joan B Hooper. 1972. The syllable in phonological theory. *Language*, pages 525–540.

Daniel Kahn. 2015. *Syllable-based generalizations in English phonology*, volume 15. Routledge.

S Prahallad Kishore and Alan W Black. 2003. Unit size in unit selection speech synthesis. In *INTERSPEECH*.

Anila Susan Kurian, Badri Narayan, Nagarajan Madasamy, Ashwin Bellur, Raghava Krishnan, G Kasthuri, MV Vinodh, Hema A Murthy, and Kishore Prahallad. 2011. Indian language screen readers and syllable based festival text-to-speech synthesis system. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 63–72. Association for Computational Linguistics.

Yannick Marchand, Connie R Adsett, and Robert I Damper. 2009. Automatic syllabification in english: A comparison of different algorithms. *Language and Speech*, 52(1):1–27.

Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

NP Narendra, K Sreenivasa Rao, Krishnendu Ghosh, Ramu Reddy Vempada, and Sudhamay Maity. 2011. Development of syllable-based text to speech synthesis system in bengali. *International journal of speech technology*, 14(3):167–181.

Ernst Pulgram. 1970. *Syllable, word, nexus, cursus*. Number 81-85. Mouton.

Kseniya Rogova, Kris Demuynck, and Dirk Van Compernolle. 2013. Automatic syllabification using segmental conditional random fields. In *BOOK OF ABSTRACTS OF THE 23RD MEETING OF COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS: CLIN 2013*, page 41. Citeseer.

Elisabeth O Selkirk. 1984. On the major class features and syllable theory. In *Language Sound Structure: Studies in Phonology*, pages 107–136. Cambridge: The MIT Press.

Leihaorambam Sarbajit Singh, Kabita Thaoroijam, and Pradip Kumar Das. 2007. Written manipuri (meiteiron) from phoneme to grapheme. *Language in India*, 7(6).

Charles Sutton and Andrew McCallum. 2006. *An introduction to conditional random fields for relational learning*, volume 2. Introduction to statistical relational learning. MIT Press.

Su-Lin Wu, ED Kingsbury, Nelson Morgan, and Steven Greenberg. 1998. Incorporating information from syllable-length time scales into automatic speech recognition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 721–724. IEEE.

Jian Zhang and Howard J Hamilton. 1997. Learning english syllabification for words. In *International Symposium on Methodologies for Intelligent Systems*, pages 177–186. Springer.