Improving the accuracy of pronunciation lexicon using Naive Bayes classifier with character n-gram as feature: for language classified pronunciation lexicon generation

Aswathy P V
Language Technology Section
CDAC-T
India
aswathypv@cdac.in

Arun Gopi
Language Technology Section
CDAC-T
India
arungopi@cdac.in

Sajini T
Language Technology Section
CDAC-T
India
sajini@cdac.in

Bhadran V K
Language Technology Section
CDAC-T
India
bhadran@cdac.in

Abstract

This paper looks at improving the accuracy of pronunciation lexicon for Malayalam by improving the quality of front end processing. Pronunciation lexicon is an in evitable component in speech research and speech applications like TTS and ASR. This paper details the work done to improve the accuracy of automatic pronunciation lexicon generator (APLG) with Naive Bayes classifier using character n-gram as feature. n-gram is used to classify Malayalam native words (MLN) and Malayalam English words (MLE). Phonotactics which is unique for Malayalam is used as the feature for classification of MLE and MLN words. Native and nonnative Malayalam words are used for generating models for the same. Testing is done on different text input collected from news domain, where MLE frequency is high.

1 Introduction

Automatic pronunciation generator is one of the main modules in speech application which determines the quality of the output. In speech applications, pronunciation is generated online from the given input using dictionary or rules. For a language like Malayalam which has agglutination, rule based approach is more suitable than look up. Since English is the official language, the influence is such that the usage of English words is very common in Indian language scripts and texts. So APLG must be able to handle the pronunciation of these English words.

The inputs given to TTS are normally bilingual with English words in Latin script and native language script. In most cases the pronunciation model for MLE is different and depends on explicit knowledge of the language. Hence, it must be identified by the system in order to enable the correct model. Language identification is often based on only written text, which creates an interesting problem. User intervention is always a possibility, but a completely automatic system would make this phase transparent and increase the usability of the system (William. B).

In this paper we brief about the language identification from text, which is typically a symbolic processing task. Language identification is done to classify MLN and MLE and apply LTS to generate Indian English pronunciation. We used Naive Bayes classifier with character n-grams as feature, to identify whether the given word belongs to native or non-native Malayalam.

2 Structure of Malayalam

Malayalam is an offshoot of the Proto-Tamil-Malayalam branch of the Proto Dravidian Language. Malayalam belongs to the southern group of Dravidian Language. There are approximately 37 million Malayalam speakers worldwide, with 33,066,392 speakers in India, as of the 2001 census of India. Basically Malayalam words are derived from Sanskrit and Tamil. Malayalam script contains 51 letters including 15 vowels and 36 consonants, which forms 576 syllabic characters, and contains two additional diacritic characters named anuswara and visarga.

In the writing system of syllabic alphabet, all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after a consonant, are used to change the inherent vowel. When they appear at the beginning of a

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 113–118, Goa, India. December 2014.

syllable, vowels are written as independent letters. When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter (omniglot). Consonants with vowel omission are used to represent the 5 chillu, which occurs in around 25% of words.

Malayalam has a canonical word order of SOV (subject–object–verb) as do other Dravidian languages. Phonetically a syllable in Malayalam consists of an obligatory nucleus which is always characterized by a vowel or diphthongal articulation; preceded and followed optionally by a 'releasing' and an 'arresting 'consonant articulation respectively (Prabodachandran, 1967 pp 39-40).

Among the vowel phonemes in Malayalam, two fronts, two are back and one is a low central vowel. Front vowels are unrounded and back vowels are rounded ones (Prabodachandran, 1967 pp 39-40). All vowels both short and long except |o| occurs initially medially and finally. Short |o| does not occur word finally medially short |e| and short |o| occur only in the first syllable. Malayalam has also borrowed the Sanskrit diphthongs of /äu/ (represented in Malayalam as ഔ, au) and /ai/ (represented in Malayalam as ഐ, ai), although these mostly occur only in Sanskrit loanwords.

In a syllable there will be at least one consonant and a vowel. By the combination of a consonant and a vowel, Malayalam syllable structure may be expressed by the simple formula (c) v (c).

The syllable structure occurring in Malayalam corpus and its frequency is shown in figure 1: C*VC* patterns occur in English words.

Harmonic sequence of vowels is one of the main characteristics of the Dravidian family of languages (Prabodhachandran).That is one vowel can influence sound of other.

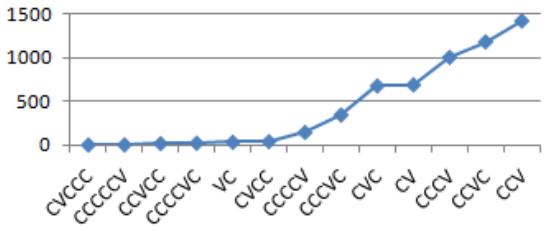

Figure 1. Syllable pattern and frequency

3 Malayalam corpus collection

The main input for speech research is a huge corpus, both text and speech properly annotated. For the major world languages, large corpora are publicly available. But for most other languages, they are not. But with the advent of the Web it can be highly automated and thereby fast and inexpensive (Adam and G V Reddy).

Malayalam corpus was created from online sources like newspapers, blogs and other sites, which is then used to extract data. Manual verification of the sites content is done to ensure that it contains good-enough-quality textual content. For domain coverage contents is manually prepared. 25GB of raw corpus is collected for extracting optimal text which covers the phonetic variations in the language. The collected corpus as such cannot be used for optimal text selection. The raw data contains junk characters, foreign words (frequent occurring of English words in Malayalam script) etc.

The optimal text selected for speech application like ASR and TTS, must be able to well represent the language. Optimal text (OT) must cover all phonemic variations occurring in the language. OT is the minimum text which covers the all/maximum possible units (variations) in the language.

Frequent occurrences of English words give higher rank for sentences containing more number of English words. This reduces the text with native content. If we have to select sentence with Malayalam words, these English words must not add to the unit count. Manually marking of these words is practically not possible. Post verification and cleaning of English words from OT is tedious and sometimes requires additional text selection.

Automatic language identification can be used in the text selection to improve the quality of OT.

4 Malayalam pronunciation and Letter to sound(LTS) rules

Malayalam is a phonetic language having a direct correspondence between symbols and sounds (Prabhodhachandran, 1997). The pronunciation lexicon generation is easy for phonetic languages when compared with English. But there exist few exceptions, in case of Malayalam.

Accurate pronunciation can be generated only by properly handling these exceptions. Pronunciation is one of the factors that determine the quality of TTS.

A detailed analysis on speech and text corpus is carried to identify the LTS rules for Malayalam. The identified rules are then verified and validated by language experts.

The Malayalam LTS rules can be categorized as below.

- Phonetic – direct correspondence between text and sounds

e.g.:
തിരുവനന്തപുരം ത് ി രു വ ന്ത പു രം
tiruvananthapuram ti ru va na ntha pu ram

- Pronunciation different from orthography

  e.g.: നനയുക ന (ന) യു ക
  "nanayuka n a n@ a y u k a" (n represent dental and n@ represent the alveolar)

  Dental and alveolar sounds for NA and its gemination

  Influence of y in gemination of kk

  Retroflex plosive and its allophone

  Lexicons having multiple pronunciations – homonym/homophones

  "ennaal" e nn aa l
  "nn" can be alveolar gemination or dental gemination. The occurrence of such cases is very less in Malayalam.

- Pronunciation by usage (Speaker specific pronunciation)

  ഉത്സവം ഉ ത്സ വം
  "utsavam" "u t s a v a m"
  -Actual pronunciation
  ഉ ൽസവം ഉ ൽസ വം
  "utsavam" "u l s a v a m"
  - by usage
  ദയ ദ യ
  "daya" "d a y a"
  – Actual pronunciation
  ദയ ദെ യ
  "daya" "d e y a" – by usage

- Foreign word pronunciation (frequent usage of English words)

For generating proper pronunciation foreign words must be identified and separate rules or pronunciation lexicon must be applied for generating proper pronunciation. The influence or dominance of English has reached to an extent that the contents from few sources even contains more than 25% of foreign words.

For handling English pronunciation 3 additional phones are required.

"a" sound as in "bank" (ബാങ്ക്)
"ph" sound as in "fan" (ഫാന്)
"s" sound as in "zoo" (സൂ)

In TTS application a standard pronunciation is used at synthesis time. Users can add user specific pronunciation dictionary to generate pronunciation in their own preference. In ASR, multiple pronunciation lexicons will be stored and pronunciation variations can be handled.

For generating accurate pronunciation, language identification is done and specific rules must be applied to handle the phone variations.

5 Pronunciation lexicon generation for Malayalam

Pronunciation lexicon for Malayalam is generated using LTS rules and exception list for handling the above listed cases.

Exception patterns for dental and alveolar NA gemination and allophonic variation of KK is extracted from corpus, and added to exception list. Pronunciation is generated based on these exception patterns. Other exceptions like pronunciation of nouns etc. are also included in this exception list.

5.1 Handling of English words pronunciation (English script)

For handling English word in latin script, encoding is checked to identify the language. Pronunciation is generated using a standard dictionary look up. CMU dictionary with 100K words is taken as the reference for English pronunciation.

A pattern based replacement is done to modify the pronunciation as Indian English (of native Malayalam speaker) pronunciation.

5.2 Handling of English word pronunciation (Native script)

The input text contains frequent occurrence of English words. Encoding based language identification is not applicable since both are in the same script.

All languages have a finite set of phonemic units on which the words are building and there are constraints on the way in which these phonemes can be arranged to form syllables. These constraints are called as phonotactics or phoneme sequence constraint. Phonotactics is language dependent.

In the process of recognition with respect to Spoken Languages it is observed that human are the best identifier of a language. However during automatic recognition we have to consider several factors like with respect to language identification one language differs from another language in one or more of the following: (M.A.Zissman, Navratil J, Mutusamy, Schultz, Mak).

- Phonology: Phone sets would be different for each of the languages
- Morphology: The word roots and the lexicons may be different for different category of languages
- Syntax: The sentence patterns with respect to grammar are different
- Prosody: Duration, pitch, and stress patterns vary from language to language

In this paper an attempt has been made to identify language using the phonology since we require word level identification. We use language specific phonotactics information to identify words that belong to native language.

5.3 Pattern based approach

Pattern based approach is a string matching method to classify words. A detailed analysis of corpus is done to extract 1200 patterns that do not occur in Malayalam. Pattern search is done on the input text and words with matching patterns were classified as English. Patterns that are common / noun patterns are not considered in the list. Only 70-75% of words are covered in this approach.

5.4 Naive Bayes classifier using character n-gram approach

Naive Bayes classifier begins by calculating the prior probability of each label, which is determined by checking the frequency of each label in the training set. The contribution from each feature is then combined with this prior probability, to arrive at a likelihood estimate for each label. The label whose likelihood estimate is the highest is then assigned to the input value.

Naive Bayes classifier with character n-gram as feature is used for categorizing MLN and MLE. In this approach two categories of texts are collected, one only with native c words (MLN) and the other with nonnative (MLE). From these texts n-gram model profile for MLN and MLE is generated. These profiles hold n-gram models up to order 3 and their individual frequencies. Generation of n-gram is shown in figure 2.

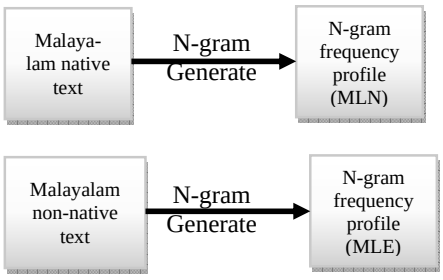

Figure 2. Generation of n-gram frequency profile for each word in the source

N-gram for Malayalam word മലയാളം

1-gram: മ ല യ ാ ള ം
ma la y aa lxa m

2-gram: _മ മല ലയ യാ ാള ളം ം_
ma mala lay yaa aalxa lxam m

3-gram: _മല മലയ ലയാ യാള ാളം ളം_
ma malay layaa yaalx aalxam lxam

As n increases expressiveness of the model increases but ability to estimate accurate parameters from sparse data decreases, which is important in categorizing nonnative Malayalam words, so n-gram up to 3 servers the job.

N-gram decomposes every string into substrings, hence any errors that are present is li-

mited to few n-grams (n-gram based text categorization, William B. Cavnar and John M. Trenkle).

5.4.1 Formulation of Naive Bayes

Given a word and the classes (C_{MLN} C_{MLE}) to which it may belong. Naïve Bayes classifier evaluates the posterior probability for which the word belongs to particular class. It then assigns the word to class with highest probability values(Jing Bai and Jian-Yun Nie. Using Language Model for Text Classification).

$$\mathbf{C_{MAP} = argmax_{c \in C} p(c|w)}$$

MAP is Maximum a posteriori probability

$$= \text{argmax}_{c \in C} \frac{p(w|c)p(c)}{p(w)}$$

$$= \text{argmax}_{c \in C} p(w|c)p(c)$$

$$= \text{argmax}_{c \in C} p(x_1, x_2, .., x_n|c)p(c)$$

x represents feature

By Naive Bayes conditional independence assumption

$$p(x_1, x_2, .., x_n|c) = p(x_1|c) \times p(x_1|c) \times ..\times p(x_n|c)$$

$$c_{NB} = \text{argmax}_{c_j \in C} p(c_j) \prod_{i \in positions} p(x_i|c_j)$$

5.4.2 Applying Naive with n-gram frequency as feature

1. Generate character based n-gram profiles for each category as discussed above
2. From training corpus extract vocabulary of model
3. Calculate $p(c_j)$ terms
   a. For each c_j in C do
      i. w_j := count of all words that belong to class C_j

$$p(c_j) = \frac{|w_j|}{W}$$

4. Calculate $p(x_i|c_j)$
   b. For each x_i

$$p(x_i|c_j) := \frac{n_{x_i} + \propto}{n + \propto |\text{vocabulary}|}$$

(Add 1 smoothing where $\propto = 1$)

5. Return C_{NB} where

$$c_{NB} = \underset{c_j \in C}{\text{argmax}} P(c_j) \prod_{i \in positions} P(x_i|c_j)$$

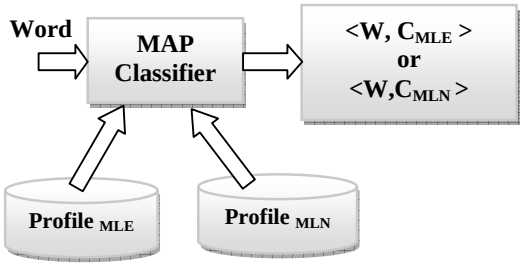

Figure 3: Categorization using Naïve Bayes

6 Experiment and results

We selected 200 sentences with 2000 unique words was taken as the input for verification of pattern based and Naive Bayes based classification. First we performed the words categorization based on the pattern list (for English words).391 words were identified and 97 were misclassified. We then used the Naive Bayes classifier to classify the words. The result is given in Table 1.0

Language identification	#Identified English word	#Miss classified
Pattern based	391	97
Naïve Bayes	782	47

Table 1.0 Category identified and misclassified

We also tested with random input collected from online sources. Naive Bayes with n-gram showed better word identification, than pattern. An average of 90% of word coverage was given by n-gram based language identification.

7 Conclusion

In this paper we brief about the effort for improving the accuracy of pronunciation lexicon for Malayalam. This method can be used to improve the quality of corpus, and speech applications like TTS and ASR. In text corpus selection, APLG using Naive Bayes classifier is used to identify foreign words in native script. This reduces the manual effort required for manual verification and cleaning of selected text corpus. For

TTS the pronunciation of words can be made accurate using Naive Bayes classifier. Depending on the domain this will increase the quality of synthetic speech.

In future we plan to improve the accuracy of Naive Bayes classifier by including the morphology rules to identify the root words. Using different smoothing technique can also improve performance(Sami Virpioja, Tommi Vatanen, Jaakko J. Vayrynen. Language Identification of Short Text Segments with N-gram Models).Naive Bayes classifier has wide range of applications which includes text categorization, development of multi-lingual speech recognizer etc.

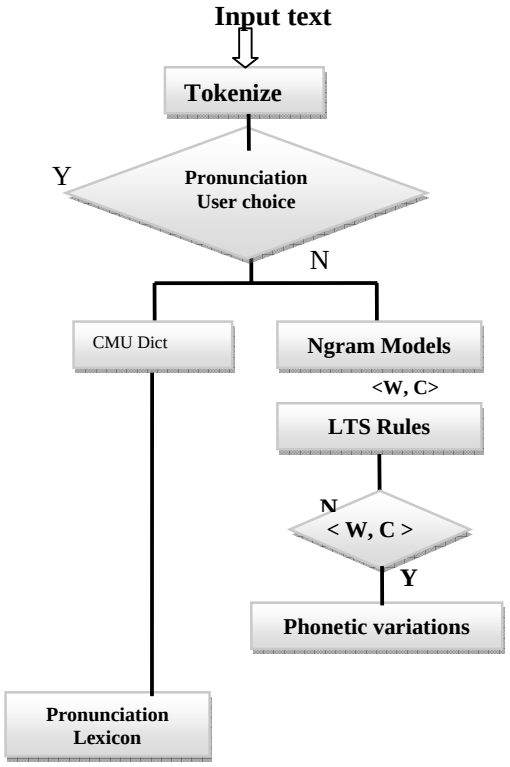

Figure 4: Implementation of APLG

Reference

Adam, G V Reddy Lexical Computing Ltd. IIIT Hyderabad, Masaryk University, IIIT Hyderabad United Kingdom, India, Czech Republic, India

M.A. Zissman, 1996 *Comparison of Four Approaches to Automatic Language Identification of Telephone speech*, IEEE Transactions on Speech and Audio Processing.

Milla Nizar, 2010, *Malayalam: Dative Subject Constructions in South-Dravidian Languages*.

Navratil. J, Sept. 2001 *Spoken Language Recognition - A Step toward Multilinguality in Speech Processing*, IEEE Transactions on Speech and Audio Processing.

T. Sajini, K. G. Sulochana, R. Ravindra Kumar, *Optimized Multi Unit Speech Database for High Quality FESTIVAL TTS*

Alan W. Black, Carolyn P. Rosé, Kishore Prahallad, Rohit Kumar, Rashmi Gangadharaiah, Sharath Rao, *Building a Better Indian English Voice using More Data*

Prabodachandran, 1967, *Malayalam structure* (Abercrombie 1 pp 39-40)

Sanghamitra Mohanty, April 2011 *Phonotactic Model for Spoken Language Identification in Indian Language Perspective*, International Journal of Computer Applications (0975 8887) Volume 19 No.9

Schultz.T, et al, 1999, *Language Independent and Language Adaptive Large Vocabulary Speech Recognition*, Proc. Euro Speech, Hungary.

Schultz. T and Kirchhoff. K, 2006, *Multilingual Speech Processing*, Academic Press

Mak. B, et al, 2002, *Multilingual Speech Recognition with Language Identification*, Proc. ICSLP.

William B. Cavnar and John M. *Trenkle N-Gram-Based Text Categorization.*

Jing Bai and Jian-Yun Nie, *Using Language Model for Text Classification*

Jaakko J. Vayrynen, Sami Virpioja, Tommi Vatanen, *Language Identification of Short Text Segments with N-gram Models*