International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 12, December-2015

Identification of Assamese and Bodo Language from Text- an Approach

Abdul Hannan
Dept. of Information Technology
Gauhati University
Guwahati, India

Shikhar Kr. Sarma
Dept. of Information Technology
Gauhati University
Guwahati, India

***Abstract*—The first step of any multilingual systems like Cross Lingual Information Retrieval and other multilingual systems is Language Identification. There are several methods have already been proposed and implemented for language identification. This paper outlines a generalized approach to language identification for text files based on techniques of rule based analysis of Assamese and Bodo language. The proposed algorithm in this paper is a three tier architecture- Unicode range checking, suffix comparison and frequent word comparison. The developed system which implements the proposed algorithm takes the plain text file as input and possesses. Although, it can also be used as a back-end tool to identify the language in online processing.**

Keywords—Language Identification, Assamese, Bodo

I. INTRODUCTION

The most important and a very basic step in almost every system related to Natural Language Processing is Language Identification. It is considered to be the first step of multilingual text processing tasks such as Summarization, Question Answering, Translation, Cross Lingual Information Retrieval (CLIR), etc. Nowadays, with the increasing use of Internet, it is becoming more usual to have the texts to be processed written in different languages. Like other languages the content of Assamese and Bodo language is also growing in electronic media. In such situation the search engines must be able to identify the languages in order to find the contents from different sources. Till date there a few reported work of language identification for Assamese language [1] and no reported work for Bodo language. In this paper we proposed one rule based approach for Language Identification based on three metric i.e. Unicode range checking, suffix checking and Frequent word checking.

II. A BRIEF OVERVIEW OF CONSIDERED LANGUAGES

A. Assamese Language

The Assamese language is cordially associated with the most important Indo-European Language Group. We have to study the Indo-European Language group to find the origin of Assamese language though it seems that it is made up of the Prota-Astrolied and Sino-Tibetan Language Group. Ascoli divided the Indo-European Language group into two main group viz. Satam and Centum. Indo-Aryan Languages are derived from the Indi-Iranian group which is one of the four sub-division of Satam. Assamese Language has also come through the three stages [(1) Old Indo-Aryan 1500BC-600BC→ (2) Middle Indo-Aryan 600BC-1000AD→ (3) New Indo-Aryan 1000AD-till now] of Indo-Aryan Language as the other Modern Indian Languages. The Indo-Aryan Languages, viz. Assamese, Bangla, Oriya etc., are derived from Avahattha through Magadhi Apravhransa. The earliest evidence of Assamese dates back to the literature of the Charyapadas, written by a few Buddhist scholars. The Assamese language present in the Charyapadas, reflects the initial stages of the evaluation of the Assamese language. There are eight vowel phonemes and twenty-one consonant phonemes including two semi-vowels in Standard Colloquial Assamese [2].

B. Morphological Characteristics

The Assamese language has many special morphological characteristics. Out of which few are outlined below:

- Numbers are not grammatically marked in Assamese Language. There are two types of Numbers in Assamese Language, Singular and Plural.
- Gender is also not grammatically marked in Assamese. Linguistically, there are two types of Gender in this Language, Masculine and Feminine. But traditionally common and neuter genders are also used.
- Kinship nouns are inflected for personal pronominal possession.
- There are two types of affixes in Assamese Language, Prefix and Suffix. Both Prefix and Suffix are very commonly found in Assamese language. Suffix included derivational, Inflectional and conjugational forms.
- There are six types of Cases in Assamese language, Nominative, Accusative, Instrumental, Dative, Ablative, and Locative. There are six parts of speech (POS) in Assamese language. They are:
  - Noun (Common, Proper, Collective, Material, Abstract.)
  - Pronoun (Personal, Demonstrative, Inclusive, Relative, Indefinite, Interrogative, Reflexive)
  - Verb (Transitive, Intransitive)
  - Adverb (Manner, Place, Time)
  - adjective (Nominal, Qualifying)

C. Syntactic Characteristics

Few general syntactic characteristics of Assamese language are mentioned below:

- The general syntactic structure of Assamese language is Subject+Object+Verb (SOV).

- Syntactically Assamese sentence structure is mainly divided into three types - Simple, Complex and Compound.
- Assamese sentence structure is flexible. Depending on the context or mood of the speaker it might vary.
- Assamese sentence structure is of different kinds. Very short sentences are found frequently. Sometimes long expressions are made by adding indeclinable.
- In Assamese, there are subject-verb agreements. The verbs in Assamese agree with the subjects in person. There is no agreement in number or gender like some other languages, English or Spanish etc.
- Verb-less sentences are also very frequent in Assamese language.

Idiomatic expressions are also found in Assamese Language.[2]

D. Bodo Language

The Bodo language has its written record from the last part of the 19th century. It was recognized by the government of Assam as official language in the Kokrajhar district and Udalguri sub-division from the year 1984. The language also got Indian government recognition as scheduled language from 2003. According to the census of 1991 it has a total of 11, 84,569 speakers.

The Bodo language belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. It is a major language of the North-Eastern part of India and has very close resemblance with the Rabha, Garo, Dimasa, Kokborok, Tiwa, Hajong and other allied languages of N-E India. It is thought that this language speakers have migrated through two different routes into Assam: one by the western route adjoining Himalayas and the other by the stream of the Brahmaputra river by eastern side of Assam. It is thought that the origin of this language is the headwaters of the Huang-Ho and Yang-Tsze-Kiang rivers in China. According to the scholars it is considered that this language presently has three distinct different dialect groups.[3]

E. Morphological characteristics

- The morphological feature of this language is discussed under two basic heads: primary and secondary grammatical categories. Primary consists of Noun, Pronoun, Verb, Adjective, Adverb, Conjunction and Interjection. Secondary consists of Number, Gender, Person, Case and Case-Endings, Numerals and Numeral Classifiers and Tense.
- Noun has basic, derived as well as compound form composed of noun and verb, verb and noun as well as noun and noun.
- Pronoun has five different categories.
- Verb has simple, complex and compound as well as transitive, intransitive, causative, finite and infinite based on structure and function.
- Adjective has basic and derived form and its basic foundation is verb.
- Adverbs are basically derived from the adjectives by using derivational suffixes.
- Numbers are two in this language and are inflected basically with nouns, pronouns also with adjectives.
- It is basically a natural gender language having two genders i.e. masculine and feminine. Traditionally common and neuter are also used. It has three different phases of gender formation.
- It has three persons: 1st, 2nd and 3rd and is discussed with the personal pronouns.
- It has seven cases including ablative and genitive.
- Numerals have basic and derived forms. Classifiers are prefixed with the numerals.
- Traditionally tense has three different forms: past, present and future, but are very difficult to completely differentiate in some cases.
- It has two affixes: prefix and suffix. In comparison to suffix the number of prefix is relatively small. Suffixes are inflectional and derivational as well as class maintaining and changing.

Kinship terms are discussed only with the personal pronouns

F. Syntactic structure

- Structurally syntax has three forms: simple, complex and compound.
- General syntactic structure is of S-O-V pattern.
- It has no concord relation.
- Its word order is flexible and is based on the context and mood of the speaker.
- It has idiomatic and non-idiomatic use of sentences.

It has the use of verb and verb less sentences.

G. Linguistic Affiliation:

Bodo belongs to the Bodo sub-section of Bodo-Naga section under the Assam-Burmese group of the Tibeto-Burman branch of the Tibeto-Chinese family.

III. Our Proposed Methodology

In our proposed approach we have considered three metric to identify the language i.e. Unicode range checking, Suffix comparison and frequent word comparison one after another.

A. Unicode Range of a Language

By checking Unicode ranges, we can identify a language group which extracts a few languages which share the same Unicode range from all the languages. In our case, if the Unicode falls between 0900 and 097F then it is from Devanagari. If the Unicode falls between 0980 and 09FF, it is from Assamese-Bengali Script.

B. Suffix list generation

A suffix is an affix which is placed after the stem of a word. Here we have considered suffix as a metric to identify the language. Since most of the languages use suffixes to derive another meaning from a root word. During the suffix list generation phase, all the valid available suffixes from each of the languages are listed first. From this suffix list only those suffixes are considered which are not valid suffix in any other language and which share the same Unicode range.

C. *Suffices of Assamese*

Some of the uncommon Assamese suffixes which are not a valid suffix for any other language which is using Assamese-Bengali script.

অৰ, টো, ঈয়া, ইলো, ইছো, ইছে, ওক, চোন, ওত

D. *Suffixes of Bodo*

Some of the uncommon Bodo suffices which are not a valid suffix for any other language which is using Devanagari script.

फोर, मोन, सोर, आव, याव, वाव, खौ

E. *Frequent word list generation*

In most of the languages there are some certain words which are used very frequently. By taking those words as parameter, a language can also be identified. Of course, there will be some common words which are frequent too. But to increase accuracy the words should be avoided.

During the frequent word generation phase, we have taken a corpus of size approximately 50,000 words for each of the languages. The corpus is generated by including content from various fields. Out of those only uncommon, frequent words are listed for comparison. There a threshold of frequency is also set as 200. i.e. Only those words are included which occurrence is at least 200 times in the random corpus.

Some of the Frequent words for Assamese

আকৌ, অথচ, অবশ্যে, বাবে, আৰু, হেতেন, যেন, হৈ, দৰে

Some of the Frequent words for Bodo

आरो, थानाय, सुलुंआव, गुबुन, थाखाय, मोनसे, गोनां, मुस्री

IV. System Architecture

We have proposed three different modules that working serially. The modules are

- Unicode range checking – This module basically checks the Unicode range of each character of a word that is fed to the system. If the range is within Assamese- Bengali Unicode range or Devanagari Unicode range then it proceeds.
- Suffix checking- This module takes the word from above module and tries to analyze whether there is a suffix in the word. If so, then checks the suffix whether it is available in the generated suffix list. If it founds it identifies the language of the word.
- Frequent word checking- This module works on word level. If the above module fails to identify, it takes the whole word and try to compare with frequent word list.

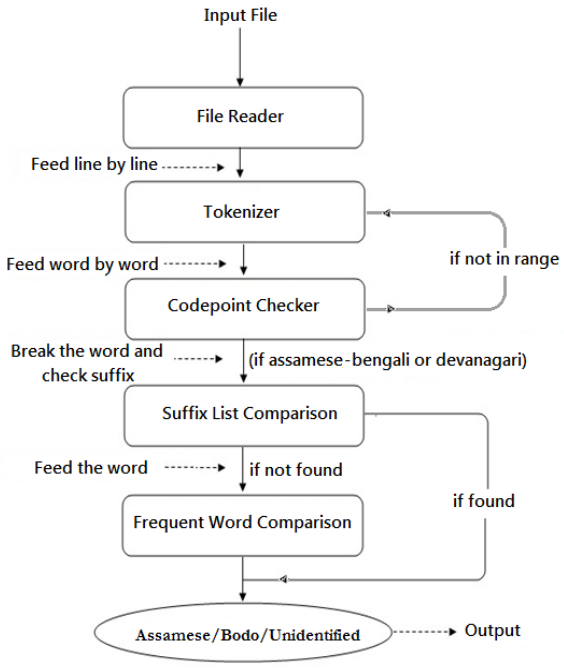

Figure 1: Modular Design of our proposed approach

V. Experimental Result

In our approach we considered different Assamese and Bodo corpuses from different sources. All the corpuses contain words from different fields such as- education, medical, sports, literature etc. We have tested with our implemented system with different size of input text also. The smallest file for experiment as input is of 10 words selected randomly. The maximum size of input file in terms of words is 7000. Here we have listed the output as correction ratio of our system.

The correction ratio is calculated as

$$P = \frac{|cw|}{|tw| - |ow| - |ud|}$$

Where,

P is the accuracy ratio for a language

|cw| is number of correctly identified words

|tw| is total number of words

|ow| is number of other language words

|uw| is number of words remained undetected

↓ Corpus size ↑		Assamese	Bodo
	~10 words	100	100
	~50 words	98.33	82.05
	~100 words	99.23	90.19
	~250 words	99.56	77.35
	~500 words	97.75	83.13
	~1000 words	99.12	82.05
	~3000 words	98.70	81.39
	~7000 words	98.55	82.07

TABLE I. CORRECTION RATIO (%) COMPARISON

CONCLUSIONS

In our work we have designed a system and implemented it to identify language of a text file. Although the system worked well for Assamese and Bodo languages with small sized text files containing a less number of words. But as the corpus size increased the accuracy of Bodo language decreased. Accuracy of Assamese language remains constant and satisfactory for all size of input. Even though 100% accuracy is not possible in case of natural language but we have got 100% in some of the cases. The minimum correction rate is 77.35%. Enhancing the rules and adding more contents to frequent and suffix list the correction ration can be increased.

REFERENCES

[1] Pinki Roy, Pradip K. Das, "International Journal of Wisdom Based Computing", 1 (3):54-59", 2011.

[2] Dr. Shikhar Kr. Sarma, R. Medhi, M. Gogoi, U. Saikia, "Foundation and Structure of Developing an Assamese Wordnet", Global Wordnet Conference, IITB, 2010.

[3] Dr. Shikhar Kr. Sarma, B. Brahma, M. Gogoi, M. B. Ramchiyari, "A Wordnet for Bodo Language: Structure and Development", Global Wordnet Conference, IITB, 2010