Language Identification System for Indian Languages

Dattesh B Naik[1] Jeevan R Patil[2] Pravin P Maske[3]

[1,2,3]Department of Computer Engineering

[1,2,3]SAOE Pune

Abstract— In the area of text classification, identification of the language is a big challenge and an important problem to solve in Natural Language Processing. When it comes to the Indian Languages the complexity increases many fold as there are cases of many languages sharing a single script and a single language written in multiple scripts. It is important to have the knowledge about the language of the text before giving it for further processing. The language identification system will be aimed towards study and research in language detection methods in practice e.g. Naïve Bayes Classifier, Random Forest Classifier, Artificial Neural Network, Support Vector Machine and the recent work happening in this area. This study and research will be used for building a system to identify the language of the input text based on the findings as a feature of classification. The language identification system will be targeted to be developed generically so that it can be adopted for identification of multiple Indian languages based on the training given to it. However, the initial emphasis will be given on the few languages especially those which are written using Devanagari script e.g. Hindi, Marathi, Konkani.

Key words: Text Classification, Language, Identification, Natural Language Processing, Indian Languages, Language Detection Methods, Naive Bayes Classifier, Markov Model, Feature of Classification, Devanagari Script, Hindi, Marathi, Konkani

I. INTRODUCTION

In the current scenario of text classification most of the researchers have focused towards different machine learning approaches. Language Identification is one of the problems in Natural Language Processing (NLP).

The Brāhmī script has been the originator of all scripts used to write Modern Indo-Aryan languages and Dravidian. Brāhmī evolved around the 3rd Century BC. Six centuries later i.e. around the 3rd Century AD, Brāhmī script in India had already divided itself into two main styles commonly termed as Northern and Southern. The northern and Southern branch of Brāhmī characterized by Indo-Aryan family and Dravidian family respectively. The Indo-Aryan, spoken by 74% Indians, and Dravidian, spoken by 24% Indians, Remaining 2% of Indians speak languages belonging to Austro-Asiatic, Tibeto-Burman and some other minor language families.

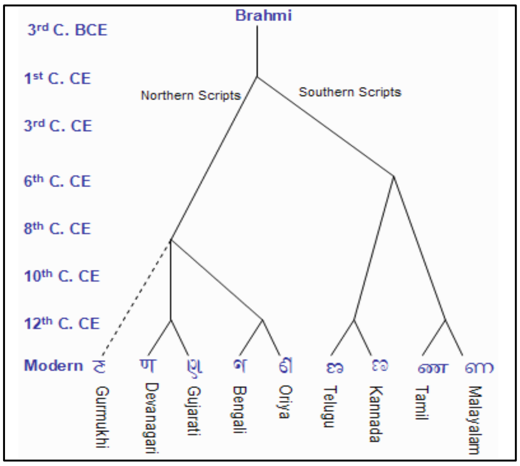

Fig. 1: Classification of Indian Languages.

India has 22 constitutionally recognized official languages and hundreds of native languages. It is difficult to identify the Indian languages as most of the languages are comes under same family of language. For solving this major problem, we are focusing on machine learning techniques. Proposed system will identify the Indian Languages by using either SVM, or ANN or Random Forest approach. Identification of language from a given small piece of text is therefore an important problem in the Indian context. Devanagari is one of the most used and adopted writing systems in the world. Devanagari script is used for writing languages like Sanskrit, Hindi, Marathi, Konkani and many other languages and dialects.

II. LITERATURE SURVEY

Lot of research has been carried out in this field and there has been significant progress in this area since last decade. Methods of language identification in practice are Naïve Based Classifier, Centric method, Support Vector Machine, Neural Networks, Markov Model, Random Forest Classifiers etc. Decision trees, Hidden Markov models, Neural Networks and SVMs are tools from more conventional pattern recognition background.

In this work we attempt to build a model for identification of languages which are derived from Devanagari script. The research we did shows that the efficiency and performance of these techniques depend on their capability in dealing with several problems such as: noisy and missing in the data used, the performance measure used, scalability, different data types used, explanation capability of the technique, ease of integration with other systems, ease of operation, and skewed distribution of the data used.

III. PROJECT SCOPE

The scope for language technologies is very high in a multilingual country like India, more linguistic initiatives would help her to emerge as a multilingual computing hub.

In this project we are going to identify the language for given input text with the help of supervised machine learning. For this we are going to work on the best approaches which are Random Forest, Support Vector Machine, Naïve Bayes and Artificial Neural Network.

The Indian language identified by proposed system will have various applications, such as

- It bridges the digital divide between the Indian people and the world.
- The linguistic initiatives would help India to emerge as a multilingual computing hub.
- Morphological processing of data.
- Spoken language identification.
- Machine translation.
- Spellchecker
- Generic system for Multiple Indian Languages.
- Discriminate languages within language families, then those across families.

IV. System Design

A. *System Architecture:*

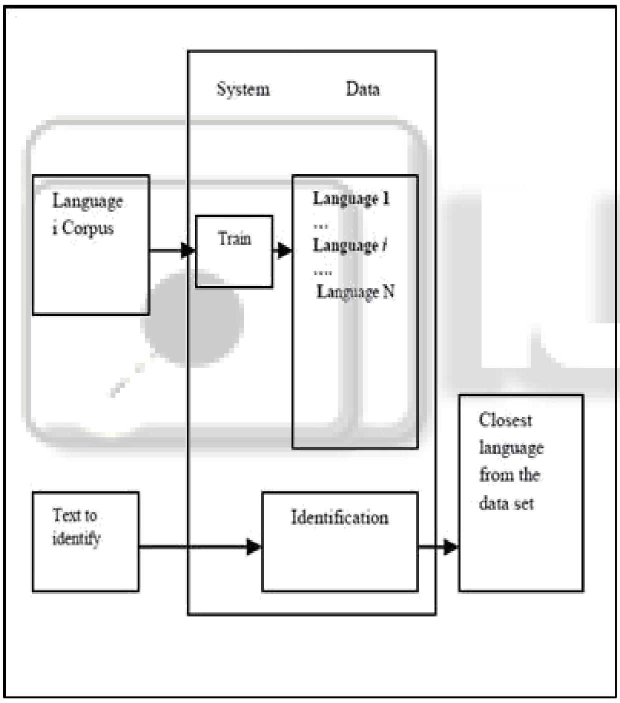

Fig. 2: General Architecture

The proposed system will work in two phases:

1) Training phase
2) Testing phase

In the Training phase, the system extracts language features from the given training corpus to generate language. A corpus is a (large) collection of electronically stored written texts. Since the languages are given, this training phase can be classified as supervised learning in the matter of machine learning.

In the Testing phase, the system classifies the language of the given document:

1) A model is generated for the given document.
2) Afterwards, the similarity between the document's model and each language models is computed.
3) The language model, which is most likely, is chosen as the language used in the document.

There are many classification approaches such as Bayesian model, decision trees, Support vector Machines, Neural Networks and Random Forests in the literature for Text Categorization.

After evaluating the above Classification Algorithms on NLTK Datasets, we have found that Random Forest Classifier is giving higher accuracy than the other classification approaches, followed by Support Vector Machine and Artificial Neural Network.

Random forest is a collection of unpruned classification or regression trees, induced from the training data, using random feature selection in the tree induction process. Each tree in the collection is formed by first selecting at random, at each node, a small group of features to split on and then by calculating the best split based on these features in the training set. To classify a new object from a feature vector, put the feature vector down each of the trees in the forest. Each tree gives a classification (votes) for that feature vector. The forest chooses the class having the most votes for a particular feature vector.

Artificial Neural Networks (ANN) have been developed as a generalization of mathematical models of biological nervous systems which is designed to mimic the decision making ability of the brain by providing a mathematical model of combination of numerous neurons connected in a network. It possesses a good learning capability that learns from given input/output data pairs and adjusts the design parameters through minimization of error function using a suitable learning algorithm.

Support vector machines are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

V. Technical Specifications

A. *Advantages:*

- Generic system for multiple indian languages.
- Discriminate languages within language families, then those across families.
- It bridges the digital divide between the indian people and the world.
- The linguistic initiatives would help india to emerge as a multilingual computing hub.

B. *Disadvantages:*

- Difficult to process non-standard words and neologisms.
- Training the model takes time on large datasets and the data should be utf-8 encoded.

  Eefficiency of language identifier depends on size of the text, to get better result input should be greater than 5 words of length.

References

[1] "langid.py", An Off-the-shelf Language Identification Tool Marco Lui and Timothy Baldwin NICTA VRL Department of Computing and Information Systems

University of Melbourne, VIC 3010, Australia mhlui@unimelb.edu.au, tb@ldwin.net

[2] "Automatic Text Categorization of Marathi Documents Using Clustering Technique" Mrs. Sushma R. Vispute Prof. M. A. Potey IEEE 978-1-4673-2818-0/13

[3] "A Comparative Study of Supervised Learning Algorithms for Re-opened Bug Prediction", Xin Xia1, David Lo, Xinyu Wang, Xiaohu Yang, Shanping Li and Jianling Sun , IEEE 1534-5351/13.

[4] "A Special Supervised Learning Algorithm and Its Applications", Yanjun Pang, Jiqiang Chen, Nianpeng Wang IEEE 978-0-7695-3745-0/09

[5] "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? " journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

[6] "Language Identification Using Wavelet Transform and Artificial Neural Network 2010 International Conference on Computational Aspects of Social Networks"

[7] "A Novel Language Identification System For Identifying Hindi, Chhattisgarhi and English Spoken Language" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERT IJERT IJERTV3IS120580 www.ijert.org Vol. 3 Issue 12, December-2014.

[8] "LANGUAGE IDENTIFICATION AND CORRECTION IN CORRUPTED TEXTS OF REGIONAL INDIAN LANGUAGES by Pooja Yadav1, Sarvjeet Kaur2 Scientific analysis Group, DRDO New Delhi, India.",

[9] "Indian Language Identification Using K-Means Clustering and Support Vector Machine" (SVM) by Vicky Kumar Verma.

[10] "Random forest algorithm for improving the performance of speech/non-speech detection", Sincy V. Thambi,Sreekumar K. T, Santhosh Kumar C, Reghu Raj P. C.

[11] "Text Based Language Identification Using Support Vector Machine", Indhuja K, Nibeesh K, P C Reghu Raj, International Journal of Computational Linguistics and Natural language Processing Vol 3 Issue 4-9 2014 ISSN 2279-0756.