Language Identification for Multilingual Machine Translation

ABSTRACT:

Machine translation is the process of translating a text in one natural language into another natural language using computer system. Translating a document containing a single source language content is easy but when the information in the source document is given in multilingual format then there is a need to identify the languages that are involved in such multilingual document. Language identification is the task in natural language processing that automatically identifies the natural language in which the content in given document are written in. Language identification is the fundamental and crucial step in many NLP applications. In this paper, n-gram based and machine learning based language identifiers are trained and used to identify three Indian languages such as Hindi, Marathi and Sanskrit present in a document given for machine translation. It is observed that, support vector machine-based language identifier is more accurate than any other technique and it achieves 89% accuracy that is 18% more than traditional n-gram based approach. The inclusion of language identification component in machine translation improved the quality of translation.

Existing System

In existing system, it is very difficult to identify language from the given text so traditionally it is very difficult by observing these issues some previous authors are using Machine learning algorithms like SVM and KNN but it is giving less accuracy and time taking process.

Disadvantages:

1. Less Accuracy
2. More time taking process

PROPOSED SYSTEM

In this project we have employed NGRAM and Machine learning algorithms to identify language names from given text. To evaluate performance, we have utilized various machine learning algorithms such as SVM, KNN and Random Forest for comparison. Each algorithm

performance is tested in terms of accuracy, precision, recall, Confusion matrix graph and FSCORE. Among all algorithms Random Forest is giving high accuracy.

Advantages:

1. High Accuracy
2. Takes less time

Modules Description:

1) **Upload Language Dataset**: using this module we will upload dataset and then remove all missing and special symbols from dataset
2) **Pre-process Dataset**: using this module we will convert above process dataset into numeric vector by employing 3 NGRAMS technique and then convert entire text data into numeric vector and then split training data into train and test where application using 80% dataset for training and 20% for testing
3) **Train KNN Algorithm**: 80% training data will be input to KNN algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
4) **Train SVM Algorithm**: 80% training data will be input to SVM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
5) **Train Random Forest Algorithm**: 80% training data will be input to Random Forest algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
6) **Comparison Graph**: will plot comparison between all algorithms
7) **Language Detection & Translation**: here user can enter some text line and then application will predict language name and then translate that language into English using Google Translator.

SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:

- Processor - Intel i3(min)
- Speed - 1.1 GHz
- RAM - 4GB (min)

- Hard Disk - 500 GB

SOFTWARE REQUIREMENTS:

- Operating System - Windows10 (min)
- Programming Language - Python (3.7.0)