JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MULTILINGUAL LANGUAGE TRANSLATOR BY DETECTING VARIOUS LANGUAGES IN THE DOCUMENT / SPEECH USING NLP

Twinkle Dharashive
Dept. of Computer Engineering
Sinhgad Academy of Engineering, Pune
twinkle055@gmail.com

Sudity Khushi
Dept. of Computer Engineering
Sinhgad Academy of Engineering, Pune
sudity7654@gmail.com

Lokesh Kolte
Dept. of Computer Engineering
Sinhgad Academy of Engineering, Pune
lokeshkolte2580@gmail.com

Rohit Mahajan
Dept. of Computer Engineering
Sinhgad Academy of Engineering, Pune
rohitm.official404@gmail.com

Abstract –

The ability to communicate with one another is a fundamental part of being human. There are nearly 7,000 different languages worldwide. As our world becomes increasingly connected, language translation provides a critical cultural and economic bridge between people from different countries and ethnic groups.

Language detection and translation is the undertaking of naturally identifying the languages present in an archive and translating them to the required language present with the records. In this work, we address the issue of distinguishing reports that contain text data present in single/multilingual records as well as an audio record that contain speech and present a technique that can recognize a record of particular language, gauge their relative meanings, and further translate them to a single language of our choice by using NLP. We exhibit the viability of our strategy over manufactured information, just as true multilingual archives would do to get a proper translation of our preferred language.

Keywords – Language Identification (LI), Support Vector Classifier (SVC), Language Modeling (LM), Natural Language Processing (NLP), Machine Translation (MT), Machine Learning.

Introduction –

Machine translation is the process of translating text from one natural language into another using computer system. You may have used google translator, which is one of famous applications of machine translation. So looking at today's world, which is full of various natural languages, it is an important need to convert the documents from a specific language to another so that people can communicate easily without any issues.

Now we have an efficient and widely used translator i.e. Google Translator, which can translate the text from a specific language into another, but what we noticed here that if we feed text that contains sentences of various languages, google translate, cannot make it. As you can see in the image.

We feed the paragraph with sentences written in Spanish, French, Hindi and English. As you can see, Google Translate detects the language as Hindi, which is not as effective as it should be and it translates just Hindi sentence to English but not others. We have tried to cover this issue and made a translator, which can detect various languages in a text and then translate it to a single language.

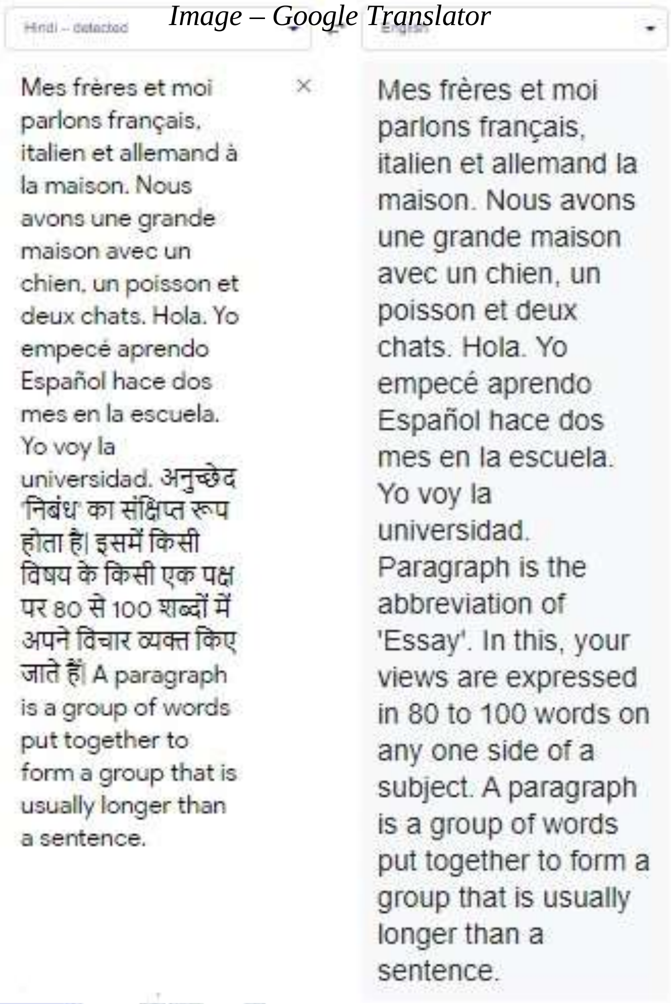
Image – Google Translator

Mes frères et moi parlons français, italien et allemand à la maison. Nous avons une grande maison avec un chien, un poisson et deux chats. Hola. Yo empecé aprendo Español hace dos mes en la escuela. Yo voy la universidad. अनुच्छेद 'निबंध' का संक्षिप्त रूप होता है। इसमें किसी विषय के किसी एक पक्ष पर 80 से 100 शब्दों में अपने विचार व्यक्त किए जाते हैं। A paragraph is a group of words put together to form a group that is usually longer than a sentence.

Mes frères et moi parlons français, italien et allemand la maison. Nous avons une grande maison avec un chien, un poisson et deux chats. Hola. Yo empecé aprendo Español hace dos mes en la escuela. Yo voy la universidad. Paragraph is the abbreviation of 'Essay'. In this, your views are expressed in 80 to 100 words on any one side of a subject. A paragraph is a group of words put together to form a group that is usually longer than a sentence.

Language identification (LI), also called as language guessing, is the task in natural language processing (NLP) that automatically identify the natural language in which the content in given document are written in. Before going for any particular natural language application one must identify the language of the content. Natural languages have different grammatical structures hence many task of NLP such as POS tagging, information extraction, machine translation, multilingual documents processing are language dependent. Language identification is fundamental and crucial stage in many NLP applications. Hence, there is need to develop an automated tool and techniques for language identification before application of further processing. For example, in case of machine translation to convert a foreign language text into required language text, the language in which the original text is written must be identified. Once it is identified then using a machine translation system it can be translated in required target language. Due to diversity of documents on the web, LI is a vital task for web search engines during crawling and indexing of web documents. For cross-lingual applications there is an increasing demand to deal with multilingual documents. Computationally, language identification problem is viewed as a special case of text categorization or classification.

This will save the cost as well as the energy. We can think how lengthy this work is going to be in bigger scale if human is going to do this task i.e. to detect the language of every sentence there in the text and then convert that sentence and so on. That is why our model can play an important role to make this task in no time.

Language Identification for Machine Translation –

Before translating the document/text, we have to identify the languages in which the text written in. Now this process is not that hard if the document/text is single lingual. Difficulty arises when this document/text becomes multilingual. We cannot process the whole text at once to the model, we may get unexpected results so we need to divide this text into segments (we may say sentences) and then then we can process these segments one by one and detect the language of each segment as shown in Fig (1).

Another factor is the dataset; we need an appropriate dataset to make this model. The records in the dataset should big enough to make the model trained and so we can get perfect accuracy. Dataset needs to be clean. We cannot process noisy dataset, which contains symbols, abbreviations, numbers, other things are alphabet cases, we can make all the alphabets into lower case as it does not make sense to use camel case words and so we need to remove all these things from the dataset as these things do not help that much in identification. We have used a dataset that contains about 10338 records in 18 languages. Each record contains a sentence and corresponding language.

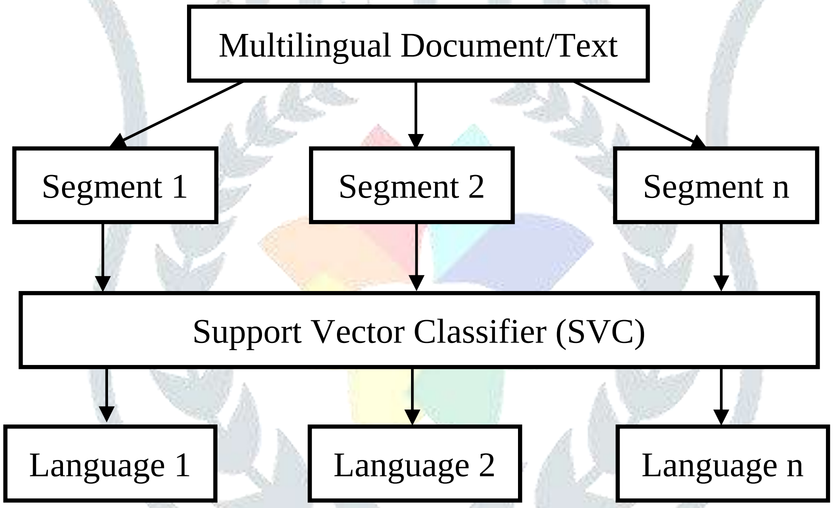

Fig 1 – Role of Language Identification in Machine Translation

Challenges involved in Language Identification –

1. Length of text data – For better results, the text data must be sufficient so that model can give accurate results. One or two word sentences can make this process even more challenging.

2. Noisy Text – Text must be clean. Use of short forms and tags known as abbreviations are considered as noisy text. It becomes challenging to decode these words, as the creator can only understand what he has written.

3. Similar words in different languages – There are some words, which are spelled similar in different languages, but their meanings are so different. The word chat means ‘an informal talk’ in English but the meaning of chat in French is a ‘Cat’ that is very different.

Machine Translation –

Now the important part i.e. machine translation of every segment should be done according to the source language and destination language. Here we have multiple source languages and one destination language, so we have to train multiple models for multiple language combinations so that it can automate all the things efficiently. Following Fig (2) explains the scenario –

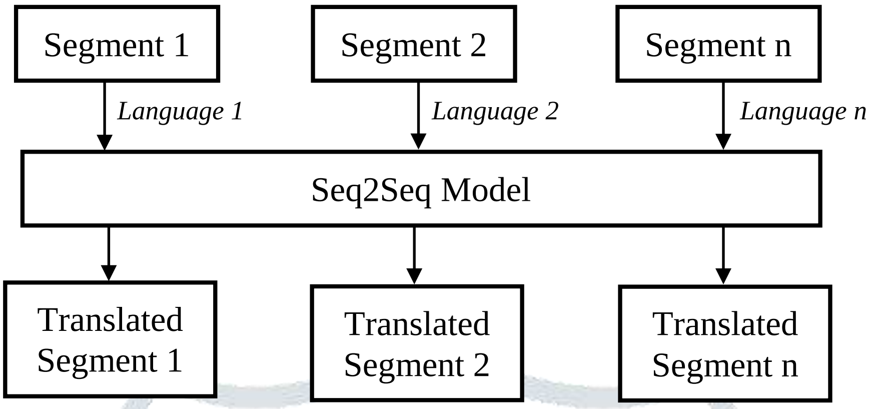

Fig 2 – Machine Translation

Methodology –

The proposed system is a Multilingual language translator where we take the input from the user and the model translates all the speech/document data into the desired text. We have used the voice recognition package "SpeechRecognition" which has a lot of hidden classes built in. One of these classes is the identity module that makes the system knows what the source is trying to say. As this audio is mp3 or some equivalent format, we use "pydub" package to convert it into wav format as to further covert the data for proper conversion into text. The converted wav audio interrupt with frequency distribution is then translated using a version that uses different APIs available in the "SpeechRecognition" package. Here we have used SVM (Support Vector Machine) to train the model for language identification. The packages like "Pipeline", "TfidVectorizer", "LinearSVC" are used to identify the language. The "TfidVectorizer" is used to vectorize the data in a format that is used to identify the language. The purpose of the "Pipeline" is to assemble several steps that can be cross-validated together while setting different parameters.

Furthermore we use Sequence to Sequence (seq2seq) model for translation. Seq2seq model is a class of Recurrent Neural Networks (RNN) that used the feed forward method to make sense of data. We use the RNN's to train the model with the help of multilingual corpus as the base for translation. The "tensorflow" package that includes sub packages like "keras" which goes further down the line from "layers" to "TextVectorization" are the main packages that are used for translation. "TextVectorization" is used for recognizing the more significant words from the less significant ones. The "keras" and "layers" package are used to encode and decode the data to translate the sentences. And this we get the result i.e. the language that the user chooses to translate.

Related Work –

The introduction of sequence to sequence model is invented by Francois Chollet. It based on model lies behind numerous systems which you face on a daily basis. seq2seq model power application such as google translate online chatbot. This application are composed of machine translation, speech recognition. There is a very large work on the application from network to machine translation. The best and effective way applying RNN language model and or a feedforward neural network to a model to MT task which improves translational quality by restoring strong best line of MT base line. Researchers have begun to look into way of finding information into the NNLM. This approach was highly successful and achieve improvements over baseline model. W. B. Cavnar and J. M. Trenkle invented n-gram for text categorization This technique is used for many researchers LI bashir ahmed improved result of document containing short strings using an ad-hoc cumulative frequency for the addition of n-grams for language identification. The accuracy of native byte classifier is more compared to rank order statistics. Bruno Martins and Mário J. For recognizing language of web document Silva developed an n-gram based algorithm worked on inverse of similarity. He implement a crawler on testing the information extracted from web page. Kosuru Pavan invented a system name ROLO to handle a random text in an Indian language like Hindi, English. The classic Soundex algorithm to deal with the phonetic distance measure for spelling differences of languages of Romanized text in Hindi, Telugu, Malayalam, Tamil and Kannada language. Abdelmalek Amine invented clustering based on unsupervised classification of multilingual text. They formed a hybrid algorithm for automatic language identification by combining artificial ant class algorithm along with n-gram based algorithm. The graph based approach proposed called LIGA to capture language grammar proposed by Erik Tromp and Mykola Pechenizkiy. They improve its accuracy with help of n-gram based algorithm for short and ill written passage in social media like Twitter, Facebook. They represent the word count in the form of graph and it also observed the ordering of word. He said that LIGA less prone to overfitting. Latin language translation successfully implemented but there are also very challenges for various Indian languages. Deepamala N, Ramakanth Kumar P give idea that use n-gram over end of sentence for identification of language like telugu, kannada. They developed character based on n gram model to obtain language from unigram to trigram. Marcos Zampieri represent word using bag-of-words approach. He uses Naïve Bayes, SVM and J48 classifier this algorithm. Kheireddine Abainia experimented two method language identification of noisy text. 1 character based method 2. term based method. Arkaitz Zubiaga uses twitter dataset and another languages. Marcos Zampieri embedded a system with the help of multiple SVM classifiers on models. Alina Maria use SVM and train model on character and word such as Hindi, Awadhi, Magadhi.

Literature Survey –

1. **Automatic Language Identification in Texts: A Survey – 2018**
   Author(s) - Tommi Jauhiainen,Krister Linden
   This article provides a brief history of LI research,and an extensive survey of the features and methods used in the LI literature.

2. **Language Identification for Multilingual Machine Translation – 2020**
   Author(s) - Arun Babhulgaonkar,Shefali Sonavane
   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.

3. **Language Detection using Convolutional Neural Network - 2020**
   Author(s) - A K M Shahariar Azad Rabby, Md. Majedul Islam, Jebun Nahar, Fuad Rahman
   In this paper, we present a lightweight, small footprint convolutional neural network, which detects Bangla and English languages—directly from scanned mixed-language document images. The proposed model achieves 99.98% recognition accuracy for this specific two-language classification problem.

4. **Identification of Languages from The Text Document Using Natural Language Processing System - 2021**
   Author(s) - *Manjula S1, Dr. Shivamurthaiah M*
   In this article One of the fundamental and significant tasks of data interpretation is language detection from textual data. The current effort is to detect the 22 distinct languages in a multilingual document using the Hybrid Isomap technique.