ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Multilingual Detection System in Indian Languages

Yugesh Sharma

M.Tech Student, Dept. of CSE, Desh Bhagat University, Mandi Gobindgarh, Punjab, India

ABSTRACT: Identification of the languages is basically the task to detect languages which is given by user to the machine or server on which it works. Language Identification detect the code of many languages likes Spanish, Dutch, Russian and many more. With the help of language identification user detect many foreign languages but still no work will be done on Indian languages. I introduce the new method of detecting various Indian Languages. This method will work on any Indian languages which is written by user or many multilingual documents.

KEYWORDS: Introduction, Literature, Survey, Issues.

I. INTRODUCTION

Identification of the languages is basically the task to detect languages which is given by user to the machine or server on which it works. Identification of languages is possible if it should be written to the close set of machine known language means the word or document which is written it should be match with the database of word or sometime close to the database word. In this method more than one Indian language will be detected and it will help to remove the monolingual assumption. I here in this paper a method is made which will detect the code of many languages which will be given by user or detect multilingual document also and estimates proportion of the word or document that is written in each languages but it should be match with database and corresponding code will be given to theuser. There are varieties of applications of multilingual detection System. Mostly Natural Language Processing process or detect monolingual input data that is why performance wise it is poor and it is also not capable of detecting more than one languages. With the help of pre-filtering step it is possible to improve the quality of input data which is use by Automatic Detection. Detecting multiple languages or multilingual document is important to separate out linguistic data from the webor sometimes for the information that how many languages are used in document and also the name of the languages like Marathi, Punjabi, Hindi and many more. With the help of this method we improve the accuracy means the result should be accurate also get the list of languages which will be written in multilingual document.
Main contribution:(1) we make a method for identifying multilingual document and number of language contained there in document.(2) It is made sure that the method will work correctly or not. The Multilingual document which will be given as a input then its corresponding output will be meet or not. (3) if there is any word which is near about database word then operation on that word will perform or not.

II. RELATED WORK

For language identification research work will basically done on monolingual document in which only single type of language will be written and that language will be detected easily. Manymultilingual document also detected but in foreign languages such as **(Marco Lui, Jey Han Lau,Timothy Baldwin)** in research paper,**("Automatic Detection and Languages Identification of Multilingual Documents")**[1] they are detect many foreign languages like English, French, Italian, German, Dutch, Japanese etc. Secondly, **(Chien Wen Yuan, Leslie D. Setlock, Dan Cosley, Susan R. Fussell)** in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

there research paper, **("Understanding Informal Communication in Multilingual Context")**[2] they are also work in many Foreign Languages but no work will be done on Indian Languages like Hindi, Punjabi, Marathi and many more. Language Identification has been applied for for finding information from the website or from online system and also in google translator such as suggestion in which the user written there textand try to translate in another language.

1.Marco Lui, Jey Han Lau and Timothy Baldwin, Automatic Detection and Language Identification of Multilingual Documents: We have presented a system for language identificationin multilingual documents using a generative mixture model inspired by supervised topic modelling algorithms, combined with a document representation based on previous research in language identificationfor monolingual documents. We showed that the system outperforms alternative approachesfrom the literature on synthetic data, as well as onreal-world data from related research on linguisticcorpus creation for low-density languages using theweb as a resource.

2. **Charles L. Wayne,Multilingual Topic Detection and Tracking:Successful Research Enabled by Corpora and Evaluation:**TDT is an important area of research, addressing central application needs. It presents new and interesting technical challenges.The enormous progress demonstrated anew the virtue of formal research task definitions, common data, and common evaluations. Clearly defined technical tasks made it possible to move forward. Representative, accurately labeled corpora made it possible to conduct meaningful research and to evaluate performance.Common, objective evaluations showed researchers which techniques worked best and allowed them to make meaningful improvements.

3. Leah S. Larkey, FangfangFeng, Margaret Connell, Victor Lavrenko, Language-specific Models in Multilingual Topic Tracking: We have confirmed the native language hypothesis for story linkdetection. For topic tracking, the picture is more complicated.When native language training stories are available, good nativelanguage topic models can be built for tracking stories in theiroriginal language. Smoothing the native models with global modelsimproves performance slightly. However, if training stories arenot available in the different languages, it is difficult to form nativemodels by adaptation or by translation of training stories,which perform better than the adapted global models. We were surprised that translating the training stories into Arabicto make Arabic topic models did not improve tracking, but again,our dictionary based translations of the topic models were differentfrom native Arabic stories. We intend to try the same experiment with manual translations of the training stories into Arabicand Mandarin.

4. Andrew G. West, Multilingual Vandalism Detection Using Language-Independent & Ex Post Facto Evidence: In this paper were motivated by changes in the 2011PAN-CLEF competition with respect to both the 2010 edition and the bulk of existing Wikipedia vandalism research. First, the competition permitted features to leverageevidence after the edits were made. We identified multiple metrics of this type, whichwere extremely effective, and whose implementation made clear the trade-off between feature efficiency and robustness.Second, the competition spanned three natural languages. For language-independent features (i.e., metadata) this was the first non-English evaluation of such signals, thoughrelative order was found to be surprisingly consistent across languages. Multiple languages,however, imply costly localization for language-specific features (e.g., profanity lists), forcing examination of their effectiveness.

5.Mr. S. B. Chaudhari, A Review on Multilingual Text to Speech Synthesis by Syllabifying the Words of Devanagari and Roman: In this system a set of words directly recorded and a set of words generated by system are played and listen by 5 listeners. Desktop speakers are used to listen these set of words. The result in table shows, each row in this table indicates the evaluation result of listeners. Result indicates listeners are in favour of syllable based proposed system; we cannot cover all languages by creating speech segment database of larger unit selection. But by using syllable unit selection speech quality is good and we can cover all languages that are uses Devanagari and roman script.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

III. OBJECTIVES

- To understand the language.
- To generate the multilingual dictionary.
- To make algorithm of multilingual system.
- To implement and test multilingual system.

IV. MULTILINGUAL DOCUMENT

In multilingual documents more than one language will be written in document and with the help of language identification we have to identify how many languages will be written and what is the code of that languages. For example document containing words in which Punjabi, Hindi, Marathi will be written then with the help of language identification we identify the words which are written in documents separately.

V. METHODOLOGY

Language Identification for multilingual document is multilevel task through these level we have to identify the languages which will be written in document as shown in above diagram 1.1.

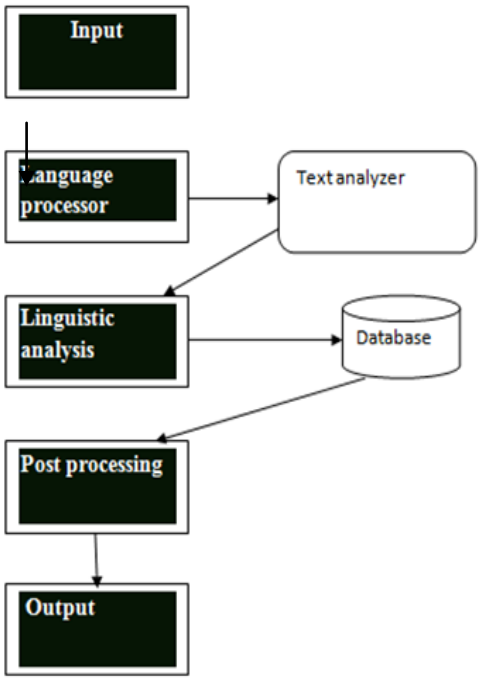

Fig 1.1 Block Diagram of Multilingual Detection System

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Firstly, the input is given to the software like words in different languages or multilingual documents. Secondly, language processor process the input words or we can say analyse that words. Thirdly, Lingustics analysis check whether the input which is given is monolingual or multilingual. Then it checks that words that it is database or not. If the words in database then post processing will be done and output will be to the user.

VI. EVALUATION

The ability of each Indian language words is evaluated.It is giving there corresponding output or not. For Example if I can type In Hindi Language then its corresponding code "hi" meet as a output similarly if I can type in Punjabi language then its corresponding o code "pa" will meet as a output. I also evaluate if I can enter more than one Indian languages words in the textbox then more than one corresponding output will meet.

VII. RESULTS

I detect all Indian languages in my project and it provide the code of that corresponding code as a output. In fig. 1.2 shown the front page of my project.

Fig 1.2 Front page of my project

When we type in Hindi Language in the textbox then its corresponding code "hi" output will meet to the user. As seen above diagram 1.3.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

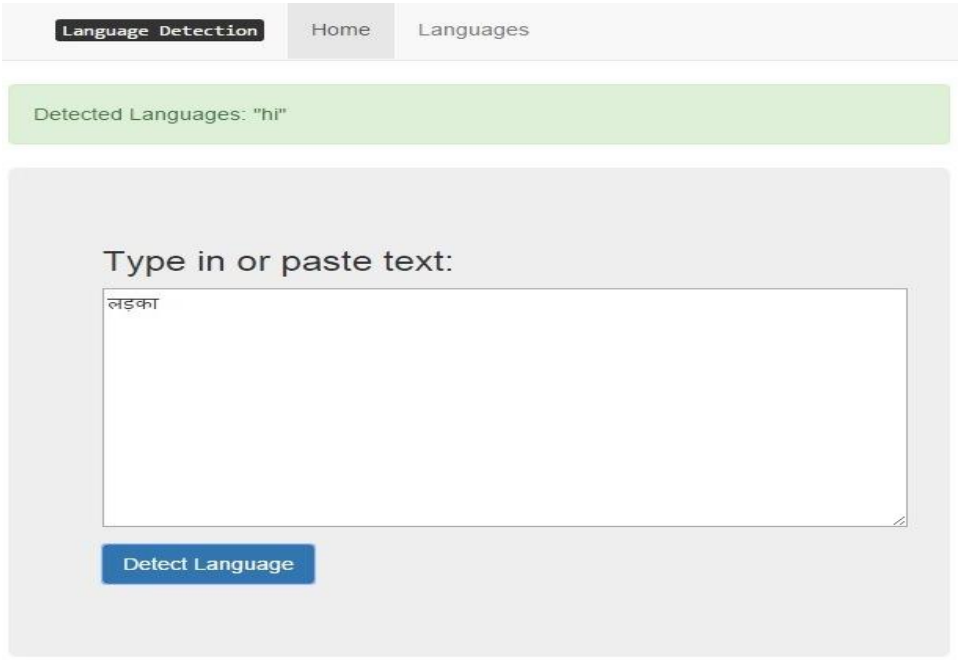

Fig. 1.3 Result of Hindi languages after detection

When we type Punjabi Languages then its corresponding code “pa” will meet to the user. As shown in above diagram 1.4

Fig 1.4 Result of Punjabi Language after detection

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

When we type in Marathi then its corresponding output code "mr" will meet to the user.As shown inbelow diagram 1.5.

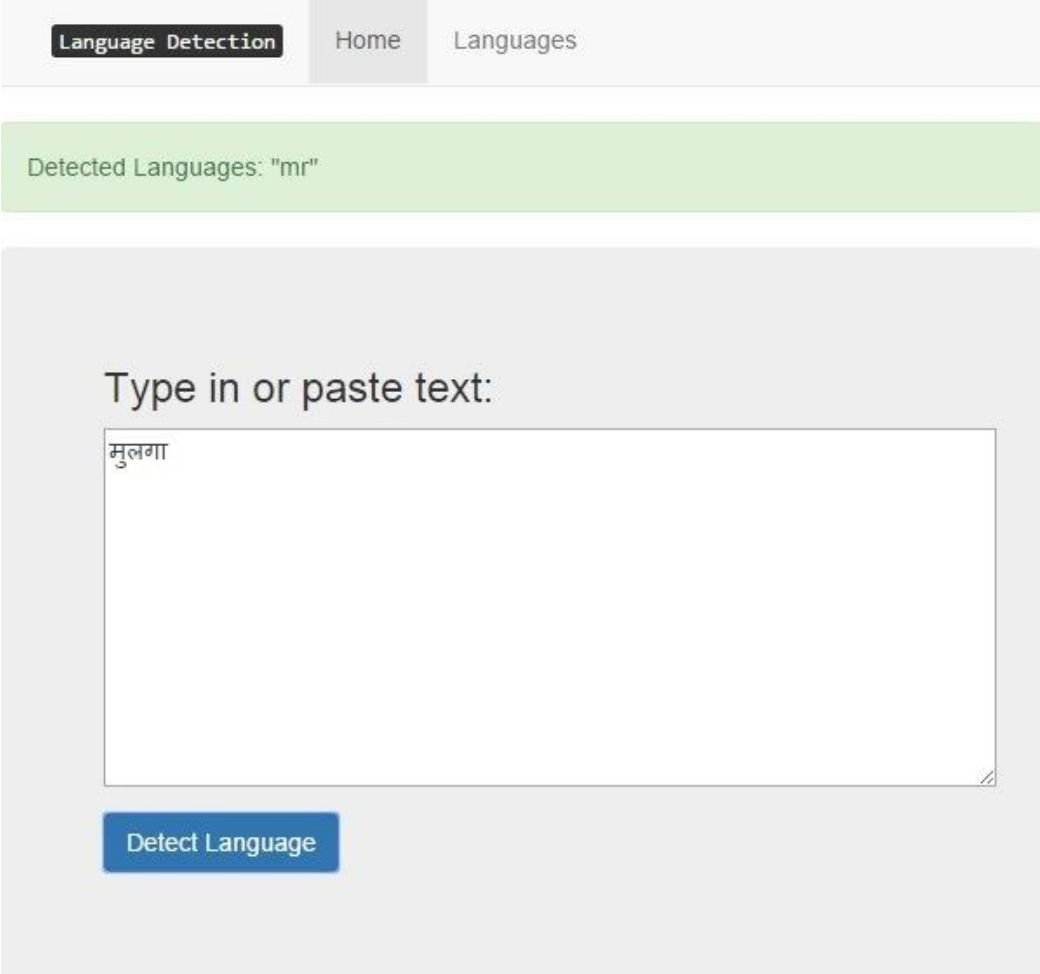

Fig 1.5 Result of Marathi language after detection

Similarly when we type all Others Indian Languages then its corresponding code as a output will meet. In this project if we type more than one languages in the textbox then its corresponding also meet to the user. Foe example in above diagram Punjabi and Telgue languages will be written and its corresponding code"pa", "te" meet to the user.As shown in fig1.6.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Fig 1.6 Result of more than one languages

VIII. CONCLUSION AND FUTURE SCOPE

After reading papers I can see that there is no work on Hindi, Punjabi, Marathi, Telgu, Malayalam, Bengali and many more Indian languages on which detection work will not be done. Then after making database of Indian languages then I detect many Indian Languages. In future I can detect the code when there is more than two languages will be written and also mention in separate way and also work on speech therapy to provide sound and corresponding codes of the language.

REFERENCES

1. Marco Lui~|, Jey Han Lau□and Timothy Baldwin,'Automatic Detection and Language Identification of Multilingual Documents, Transactions of the Association for Computational Linguistics, 2 (2014) 27–40. Action Editor: Kristina Toutanova.Submitted 1/2013; Revised 7/2013; Published 2/2014. c2014 Association for Computational Linguistics'.
2. Chien Wen Yuan, Leslie D. Setlock, Dan Cosley, Susan R. Fussell,'Understanding Informal Communication in Multilingual Contexts,'*CSCW '13,* February 23–27, 2013, San Antonio, Texas, USA.Copyright 2013 ACM 978-1-4503-1331-5/13/02...$15.00.
3. Leah S. Larkey, FangfangFeng, Margaret Connell, Victor Lavrenko,'Language-specific Models in Multilingual Topic Tracking,'*SIGIR '04,* July 25-29, 2003, Sheffield, South Yorkshire, UK.
Copyright 2004 ACM 1-58113-881-4/04/0007…$5.00'.
4. Yi Chang, Ruiqiang Zhang, Srihari Reddy,'Detecting Multilingual and Multi-Regional Query Intent in Web Search, Copyright _c 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.'.
5. Andrew G. West, Insup Lee" Multilingual Vandalism Detection Using Language-Independent & Ex Post Facto Evidence,'Notebook Papers on Uncovering Plagiarism, Authorship, and Social Software Misuse, Amsterdam, the Netherlands. September 2011.
6. Dr. Matthias Hecking, Dr. Andreas Wotzlaw, Ravi Coote,'Multilingual Content Extraction Extended with Background Knowledge for Military Intelligence,'16th ICCRTS - International Command and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Control Research and Technology Symposium'.
7. Aleksander M. Stensby,"Language Detection and Tracking in Multilingual Documents Using Weak Estimators'.
8. Martijn Spitters,Wessel Kraaij,"Unsupervised Event Clustering in Multilingual News Streams'.
9. Yuling Pan, Jennifer Leeman,Marissa Fond,Patricia Goerman,"Multilingual Survey Design and Fielding:Research Perspectives from the U.S Census Bureau'.
10. Georgiana Pu¸sca¸su,'A Multilingual Method for Clause Splitting'.
11. https://en.wikipedia.org/wiki/Languages_of_India.