Text Based Language Identification System for Indian Languages Following Devanagiri Script

Indhuja K, Indu M, Sreejith C
M.Tech Computational Linguistics
Department of Computer Science and Engineering
Government Engineering College
Sreekrishnapuram, Palakkad,
India

P. C. Reghu Raj
Professor
Department of Computer Science and Engineering
Government Engineering College
Sreekrishnapuram, Palakkad,
India

***Abstract*—: Text based language identification is the task of automatically recognizing a language from a given text of document. It is difficult to discriminate languages within language families than those across families. In this paper, we investigate the performance of statistical measures to determine the text-based language identification system, with an emphasis on five languages used in India based on Devanagiri script - Hindi, Sanskrit, Marathi, Nepali and Bhojpuri. The proposed system uses n-grams as feature for classification. Language Identification is an important pre-processing step in many tasks of Natural Language Processing (NLP). In a multilingual society like India there is wide scope for automatic language identification since it would be a vital step in bridging the digital divide between the Indian masses and the world.**

Keywords— Devanagiri Script, Multilingual Computing, Natural Language Processing, n-gram Statistics, Text Based Language Identification

I. INTRODUCTION

Language identification (LID) is an important problem in the field of Natural Language Processing (NLP). With the current spread of internet, text is available in number of languages other than English. The automatic treatment of these texts, for any purpose requiring NLP, such as indexing, interrogation necessitates the primary identification of language. It may seem to be an elementary and simple issue for humans in the real world, but it is difficult for a machine, primarily because different scripts are made up of different shaped patterns to produce different character sets.

LID is of special significance especially for multi-lingual country like India. There are a large number of languages used in India, of which twenty two have been given constitutional recognition and are considered major languages [14]. In most cases, frequent code switching and code mixing are also observed. If we could segment multi-lingual documents language-wise, it would be very useful both for exploration of linguistic phenomena, such as code-switching and code mixing, and for computational processing of each segment appropriately. 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, Nepali, Konkani, Punjabi and many other languages and dialects.

One of the popular methods for language identification is the n-gram based method. n-gram method uses letter n-grams representing the frequency of occurrence of various n-letter combinations in a particular language. In n-gram based methods for text based LID, frequency statistics of n-gram occurrence are used as features in classification. The advantage of using n-gram over other methods is that no linguistic knowledge needs to be gathered to construct a classifier. n-gram methods are simple, the accuracy increases with the increasing size of n . Long character strings contain more n-grams and statistical measures can be calculated from it. The number of n-grams in a character string is equal to $l-n+1$, where l is the length of string.

Our objective is to build a text based language identification system for Indian languages following Devanagiri script. This paper is organized as follows. Section 2 describes detailed literature survey that helps to formulate the problem. In Section 3 an n-gram model proposed for identifying the given language pairs. The experimental details and the results obtained are presented in section 4. Conclusions are given in section 5. Last section includes references.

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 Naive Based Classifier, Centric method, Support Vector Machine, Neural Networks, Markov Model etc. Here we discuss some recent studies carried in the field of language identification. Decision trees, Hidden Markov models, Neural Networks and SVMs are tools from more conventional pattern recognition background. Though it may be expected that these classifiers would prove more accurate in the task, published results demonstrate that it is still difficult to outperform the simpler methods. In n-gram based methods for text-based LID,

frequency statistics of n-gram occurrences are used as features in classification [1].

Gerrit Reinier Botha's work on language identification for the South African languages uses n-gram statistics for classification and compared with different text based language identification approaches[1]. In the paper "Using n-grams for LID", Combrinck and Botha presented a text-based language identification system for 12 languages, including six African and six European languages. A crucial part of the system was the identification of the set of most distinctive, most frequently encountered sequences of characters (i.e. ngrams) that could be associated with each language. On the basis of frequency of occurrence, the score was assigned and based on the score the text was classified. Tomas Olverky, discusses the ability of n-gram categorization to classify an unknown text. This paper focused on how n-grams could be tuned to perform better and proposed methods for improvement [3]. Grigory Grefenstette analyzed two techniques - trigrams and short words for language identification and found that trigrams perform better for small sentences while bigram work well for longer sentence [4]. A paper by Aditya Bhargava et.al shows an approach based on SVMs in LID with n-gram counts as features [5]. This papers shows n-gram model outperforms all other models in language modeling. The paper by Majumder, M Mitra, " N-gram: a language independent approach to IR and NLP ", explains the significance of n-gram in the field of NLP and language modeling [6] . Tommi Vatanen et.al, compares two distinct methods that are well suited for LID task: a Naive Bayes Classifier based on character n-gram models, and the ranking method by Cavnar and Trenkle . The accuracy of the studied methods was found to decrease significantly when the identified text gets shorter [7]. These works instinct us to implement our work on Language identification of Indian languages based on n-gram model.

Some of works on LID in Indian languages are remarkable and these works helps us to know challenges and methods of Indian language identification. Kavi Narayana Murthy formulated language identification as machine learning problem, a supervised classification task in which features extracted from a training corpus are used for classification [8]. The paper Using n-gram and Word Network Features for Native Language Identification, by Shibamouli Lahiri identifies writer's native language from his/ her writing in second language using n-gram feature and WordNet[9]. Another method in LID of Indian languages proposed by Pinky Roy's as "Language Identification using Gaussian Mixture Model Tokenization", aims at identifying the language of a spoken utterance. It uses Gaussian mixture model as basis phone tokenization and uses n-gram for identification [10].

Hindi is one of the official languages of India and is the native language of people living in Delhi, Haryana, Uttar Pradesh, Bihar, Jharkhand, Madhya Pradesh and parts of Rajasthan [16]. Sanskrit one of the 22 scheduled languages of India and is an official language of the state of Uttarakhand [15]. Bhojpuri is a North Indian language and is used in Bhojpuri region of North India and Nepal. It is spoken in the Purvanchal region of Uttar Pradesh, in the western part of state of Bihar, and the northwestern part of Jharkhand in India [4]. Nepali or Nepalese is a language in the Indo-Aryan language family. It is the official language of Nepal and is also spoken in Bhutan.

In this work we attempt to build a model for identification of Indian languages which use Devanagiri script. We make use of frequent words and character statistics for language identification on basis of n-gram along with fundamental property of character frequencies to identify Indian languages. We have used n-gram statistics for language identification among Indian languages including Hindi, Sanskrit, Marathi, Bhojpuri and Nepali.

III. PROPOSED SYSTEM

Recent research on language identification has been limited exclusively to machine learning approaches. In machine learning approaches, a set of training data is given and the machine ''learns'' a general rule or builds a model for performing the intended task. A machine learning system is expected to be generic and it is understood that training is based only on the intrinsic properties of the data, as expressed through a set of ''features''.

As discussed above, we restrict our attention to the Devanagari languages – in particular, the five languages used in India. Text from various domains in all five languages was obtained from various sources such as newspapers, periodicals, books etc. Therefore, the extracted corpus spans several domains. For this, initially, separate corpora of approximately 2MB size were created for Hindi, Nepali, Bhojpuri, Marathi and Sanskrit languages. This was done by extracting text from Wikipedia and other online documents. Then the language corpus was filtered and sent as parameter to the training profile generator to generate character-based as well as word-based unigram, bigram and trigram training sets for all text data. Thus, separate language profiles for languages were formulated based on n-gram frequency from the corpus.

Natural language Tool Kit (NLTK) in Python language was used for this experimentation. Once the training set was created, the proposed system was used on random test data for classification and identification of unknown content in the digital online text. This test data was taken randomly from Internet with sentences in any of these five languages and the results were noted. Then, the test data was filtered and sent as parameter to the testing profile generator code to generate character based as well as word based unigram, bigram and trigram training set text data. Then the similarity measures of languages were calculated. The flow diagram is given below:

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one

return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

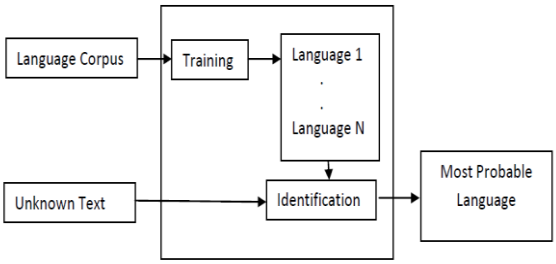

Fig.1 Steps to identify a particular language.

During training stage, all possible n-grams (unigram, bigram and trigram), both character level and word level were extracted. The core advantages of n-gram models (and algorithms that use them) are relative simplicity and the ability to scale up by simply increasing n. The model can be used to store more contexts with a well understood space–time tradeoff, enabling small experiments to scale up very efficiently. The n-gram approximation for calculating the next word in the sequence is given by:

$$P(X_{1.....}X_n) = P(X_1)\,P(X_2|X_1)..... P(X_n|X_1^{n-1})$$

$$= \prod_{k=1}^{n} P(X_1^{k-1})$$

LID system has two components - Language Profile Generator and Classifier. For language identification, the former calculates the n-gram profile of a text to be identified and compares it to language specific n-gram profiles. For each language, it generates all possible n-grams for the text and save it into corresponding language files. In classification method, given a test sample, its likelihood is calculated for all the models, and the language that gives the best likelihood is selected.

TRAINING

ALGORITHM: LANGUAGE PROFILE GENERATION
Input : Language corpus of Hindi, Sanskrit, Nepali, Bhojpuri, Marathi.

Output: n-grams of these languages

1. The documents from corpus are taken one by one, and the preprocessing is done i.e. removal of special characters, digits etc
2. Tokenize the text in to words (tokens).
3. Generate all possible character and word based n-grams (for n =1 to 3) for each language.
5. Sort n-grams based on their frequencies using frequency distribution.
6. Store n-gram profiles of all language sets (classifier).

TESTING

ALGORITHM:
LANGUAGE IDENTIFICATION

Input : Text in Hindi, Nepali, Marathi, Bhojpuri or Sanskrit.

Output: Language identified for the given text.

1. Read the text to be identified from user.
2. Remove the special characters and tokenize the text into tokens.
3. Generate all possible n-grams for the text.
4. Save it into corresponding files of languages.
5. Calculate similarity index of given languages by comparing it with the result of language profile generators.
6. Select the language corresponding with the highest similarity value.

Most Indo-Aryan languages are written in Devanagari scipt . Devanagari is an alpha syllabary and the heart of the writing system is the syllable or akshara. An alpha syllabary is a writing system which is primarily based on consonants, and in which vowel symbols are requisite [8]. The Unicode Standard defines three blocks for Devanagari : Devanagari (U+0900–U+097F), Devanagari Extended (U+1CD0–U+1CFF), and Vedic Extensions (U+A8E0–U+A8FF). Grey areas indicate non-assigned code points. The range is from 0900 to 097F.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+090x	ऀ	ँ	ं	ः	ऄ	अ	आ	इ	ई	उ	ऊ	ऋ	ऌ	ऍ	ऎ	ए
U+091x	ऐ	ऑ	ऒ	ओ	औ	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट
U+092x	ठ	ड	ढ	ण	त	थ	द	ध	न	ऩ	प	फ	ब	भ	म	य
U+093x	र	ऱ	ल	ळ	ऴ	व	श	ष	स	ह	ऺ	ऻ	़	ऽ	ा	ि
U+094x	ी	ु	ू	ृ	ॄ	ॅ	े	ै	ॉ	ॊ	ो	ौ	्	ॎ	ॏ	ॐ
U+095x	॑	॒	॓	॔	ॕ	ॖ	ॗ	क़	ख़	ग़	ज़	ड़	ढ़	फ़	य़	
U+096x	ॠ	ॡ	ॢ	ॣ	।	॥	०	१	२	३	४	५	६	७	८	९
U+097x	॰	ॱ	ॲ	ॻ	ॼ	ॽ	ॾ	ॿ		ঀ	ঁ	ং	ঃ	঄	অ	আ

Fig.2 Devanagari Unicode chart (As of Unicode version 6.3)

IV. RESULT AND EVALUATION

The proposed system was trained using a corpus of languages like Hindi, Sanskrit, Marathi, Bhojpuri and Nepali having size 2MB. The proposed LID uses two level monograms, bigrams and trigrams. The most frequently occurring differential features are retained. Feature extraction

is done using the corpus i.e. character level monograms, bigrams and trigrams that occur frequently in the language are found out. In the next level, the proposed system was tested using a data set of 100 samples each. Samples are extracted randomly from web. Each sample was analyzed into characters and the differential feature values were obtained. The pair wise result analysis of languages is done for the sample. Input of two level LID is text of any size greater than 5 words in any of these 5 languages i.e. Hindi, Nepali, Sanskrit, Marathi, Bhojpuri. Output will be the language identified for the given text. Here we explore all possible combination of languages. The accuracy of 2 pair, 3pair, 4 pair and 5 pair is computed on the basis of n-gram features. The accuracy of different pairs are as shown in the table.

Language	Unichar	Bichar	Trichar	Uniword	Biword	Triword
Hindi - Bhojpuri	60	80	72	**90**	60	32
Hindi - Marathi	74	**90**	81	84	36	20
Hindi - Nepali	56	**92**	82	90	54	24
Hindi - Sanskrit	78	96	92	**100**	32	8
Bhojpuri – Marathi	76	96	92	**100**	26	3
Nepali - Marathi	86	98	98	90	44	2
Bhojpuri -Nepali	85	98	**100**	80	44	6
Bhojpuri -Sanskrit	78	88	**96**	86	22	0
Nepali- Sanskrit	78	88	**96**	86	24	0
Marathi- Sanskrit	86	**100**	**100**	82	34	0

Table 1: Testing with two languages (Accuracy in Percentage)

Language	Unichar	Bichar	Trichar	Uniword	Biword	Triword
Hindi Bhojpuri Nepali	44	73	70	**80**	77	45
Hindi Marathi Sanskrit	64	93	82	81	32	24
Hindi Nepali Sanskrit	57	94	78	84	52	1.3
Marathi Hindi Bhojpuri	68	73	**78**	**78**	41	20
Nepali Hindi Marathi	62	94	94	**96**	24	26
Sanskrit Nepali Marathi	60	**90**	80	77	18	0
Bhojpuri Hindi Marathi	68	76	**78**	61	36	14
Nepali Bhojpuri Marathi	42	94	94	**96**	24	26
Marathi Bhojpuri Sanskrit	70	96	**98**	80	26	10

Table 2: Testing with three languages (Accuracy in Percentage)

Language	Unichar	Bichar	Trichar	Uniword	Biword	Triword
Hindi Marathi Nepali Bhojpuri Sanskrit	56	72	82	88	38	8

Table 4: Testing with five languages (Accuracy in Percentage)

Language	Unichar	Bichar	Trichar	Uniword	Biword	Triword
Nepali Bhojpuri Marathi Hindi	53	93	88	97	42.5	2.5
Hindi Sanskrit Bhojpuri Marathi	57.5	80	95	100	85	27.5
Nepali Sanskrit Hindi Bhojpuri	65	75	82	85	47	25
Nepali Sanskrit Hindi Marathi	65	90	85	90	20	0
Sanskrit Marathi Bhojpuri Nepali	70	95	92	95	42.5	2.5

Table 3: Testing with four languages (Accuracy in Percentage)

From table 1, it can be concluded that trichar, uniword and bichar identify the corresponding language with a greater accuracy around 90% in two pair LID. It is observed that the differences between the languages within language family and across language family cases are not very drastic. For example Hindi is inherited from Sanskrit, similarly Marathi, Bhojpuri are from Hindi. They show a narrow gap in language identification. Table 2 shows that accuracy decreases when going from two pair to three pair. The interdependency of language downs the accuracy level. The language that inherited from same ancestor shows much similarity in between them than the ancestors. Average accuracy decreases from 87.2 to 83.5. Here higher accuracy is shown by Nepali-Hindi-Marathi and Marathi-Bhojpuri-Sanskrit. It is noticed that the lexical similarity between these languages are less. Table 3 shows that four pair accuracy is better Sanskrit -Marathi-Nepali-Bhojpuri. As these languages, has no interconnection in between them. The common language in the inheritance hierarchy is Hindi. Accuracy is low for Nepali-Hindi-Bhojpuri-Sanskrit as they are interrelated in inheritance. Lexical similarity is too high for these languages. The average accuracy in 5 pair is 80% much less than other pairs.

V. CONCLUSION

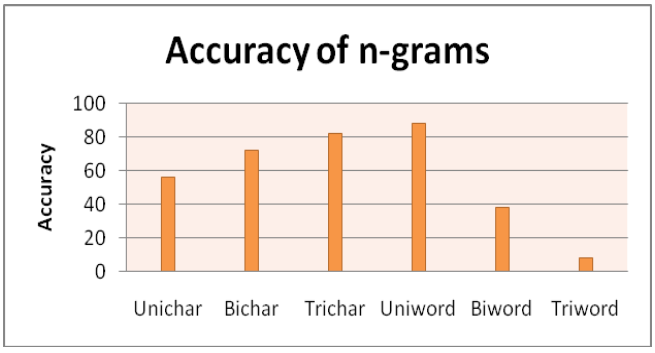

Fig.3. n-gram classifier accuracy for five language

The following are the Inference from the LID of Indian language. The words that occur mostly in one language may not occur in another. Though all languages have similar script, the meaning of words formed from aksharas changes. The position of characters (aksharas) or words changes from one language to another. The ending case marker (words or characters) changes from one language to another. Bigrams that are frequent in one language may not be same as another language. Similarly the occurrence of trigram in one language may not be same as another language. As the data set for testing increases, then the accuracy of result also increases. Lexical similarity between languages hinders the Language identification task. Greater lexical similarity between languages, accuracy of the LID will be less. The problems faced in LID of Indian languages are:-

- In Indian languages, several words are derived from single root word. These words have the root form common in all languages.
- It is not feasible to collect all the words in any Indian language to form a dictionary and search in the dictionary.
- Efficiency of Language identifier depends on size of the text, to get better result input should be greater than 5 words of length.
- Lack any explicit representation of long range dependency

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.

REFERENCES

1. Gerrit Reinier Botha, "Text based language identification for the South African Languages". Master's Dissertation, Department of electronic and computer Engineering ,April 9th 2008.
2. Combrinck, H., & Botha. E, "Automatic language identification: Resisting Complexity ". South African Computer Journal, 27, 18 – 26, 1995
3. Tomáš Olverky, "N-Gram Based Statistics Aimed at Language Identification", Slovak University of Technology, M. Bieliková (Ed.), IIT.SRC , , pp. 1-7, April 27, 2005
4. Gregory Grefenstte "Comparing two language identification schemes", 3rd International conference on Statistical Analysis of Textual Data, Dec 11–13, 1999
5. Aditya Bhargava and Grzegorz Kondrak, "Language identification of names with SVMs". Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 693–696, 2010
6. P Majumder, M Mitra, B.B. Chaudhuri ," N-gram: a language independent approach to IR and NLP" , International Conference on Universal Conference on Universal Knowledge and Language, 2002
7. Tommi Vatanen, Jaakko J. Vayrynen, Sami Virpioja, "Language Identification of Short Text Segments with N-gram Models", European Language Resources Association, 2010
8. Kavi Narayana Murthy and G. Bharadwaja Kumar , "Language Identification from Small Text Samples", Journal of Quantitative Linguistics, 2006
9. Shibamouli Lahiri, Rada Mihalcea, " Using N-gram and Word Network Features for Native Language Identification" , Proceedings of the The 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications , June 2013.
10. Pinki Roy, Pradip K. Das, "Language Identification of Indian Languages Based on Gaussian Mixture Models", International Journal of Wisdom Based Computing, Vol. 1, December 2011
11. Simon Kranig, "Evaluation of Language, Identification Methods." University of T'ubingen, International Studies in Computational Linguistics.
12. B. Ahmed, S. Cha, "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", Proceedings of CSIS 2004, Pace University, May 7th, 2004
13. Sreejith C, Indu M, P. C. Reghu Raj, "N-gram based Algorithm for distinguishing between Hindi and Sanskrit texts", IEEE Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013
14. Languages in India – http://www.mapsofindia.com/culture/indian-language.html
15. "Sanskrit is second official language in Uttarakhand – The Hindustan Times". Hindustantimes.com. 19 January 2010.
16. Hindi. Keith Brown, ed. Encyclopedia of Language and Linguistics (2 ed.). Elsevier. ISBN 0-08-044299-4, 2005