Statistical Analysis of Multilingual Text Corpus and Development of Language Models

Shyam S. Agrawal, Abhimanue, Shweta Bansal, Minakshi Mahajan

KIIT College of Engineering, Gurgaon, India

dr.shyamsagrawal@gmail.com, er.abhimanue@gmail.com, bansalshwe@gmail.com, minakshimahajan@gmail.com

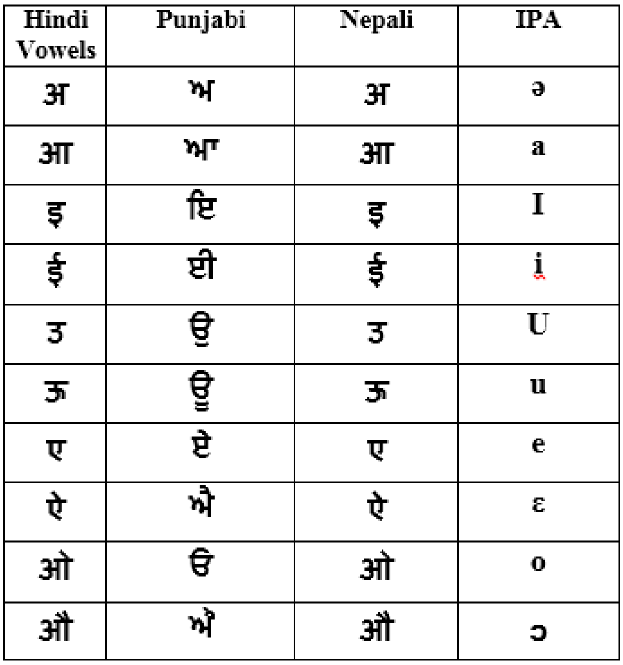
Abstract

This paper presents two studies, first a statistical analysis for three languages i.e. Hindi, Punjabi and Nepali and the other, development of language models for three Indian languages i.e. Indian English, Punjabi and Nepali. The main objective of this study is to find distinction among these languages and development of language models for their identification. Detailed statistical analysis have been done to compute the information about entropy, perplexity, vocabulary growth rate etc. Based on statistical features a comparative analysis has been done to find the similarities and differences among these languages. Subsequently an effort has been made to develop a trigram model of Indian English, Punjabi and Nepali. A corpus of 500000 words of each language has been collected and used to develop their models (unigram, bigram and trigram models). The models have been tried in two different databases- Parallel corpora of French and English and Non-parallel corpora of Indian English, Punjabi and Nepali. In the second case, the performance of the model is comparable. Usage of JAVA platform has provided a special effect for dealing with a very large database with high computational speed. Furthermore various enhancive concepts like Smoothing, Discounting, Backoff, and Interpolation have been included for the designing of an effective model. The results obtained from this experiment have been described. The information can be useful for development of Automatic Speech Language Identification System.

Keyword–Statistical Analysis, Trigram Model, Multilingual Corpus, Language Models

1. Introduction

In India people speak a variety of languages and each language has several dialects. According to present linguistically based classification, the official languages have been classified into 22 languages and more than 300 dialects spoken in different parts of India. In the present paper the languages considered for statistical study include Hindi, Punjabi and Nepali and for developing language model Indian English, Punjabi and Nepali have been chosen. Hindi is spoken by maximum number of people by about 41% of the population mostly in northern, central, western and eastern parts of the country. Punjabi is spoken by about 2.8% population, Nepali by about 0.3% mostly in northeast areas and English (which can be called as Indian English) by about 3 to 5% of the total population distributed in different parts of India[1]. The reliability and coverage of the optimal text and of the language model largely depends on the quality of the text corpus chosen. The corpus should be unbiased and large enough to convey the entire syntactic behaviour of the language. In the present experiment, 500,000 words for each language which were collected from various domains of general and special purpose communication. Table 1(a) shows the vowels and Table 1(b) consonants along with their transcription symbols used to represent the phonemes of these languages.



Hindi Vowels	Punjabi	Nepali	IPA
अ	ਅ	अ	ə
आ	ਆ	आ	a
इ	ਇ	इ	I
ई	ਈ	ई	i
उ	ਉ	उ	U
ऊ	ਊ	ऊ	u
ए	ਏ	ए	e
ऐ	ਐ	ऐ	ɛ
ओ	ਓ	ओ	o
औ	ਔ	औ	ɔ

Table1(a): Pure Vowels of Hindi, Punjabi and Nepali with their IPA transcription symbols.

Hindi	Punjabi	Nepali	IPA
क	ਕ	क	k
ख	ਖ	ख	kʰ
ग	ਗ	ग	g
घ	ਘ	घ	gʱ
ङ	ਙ	ङ	ŋ
च	ਚ	च	tʃ
छ	ਛ	छ	tʃʰ
ज	ਜ	ज	dʒ
झ	ਝ	झ	dʒʱ
ञ	ਞ	ञ	ɲ
ट	ਟ	ट	ʈ
ठ	ਠ	ठ	ʈʰ
ड	ਡ	ड	ɖ
ढ	ਢ	ढ	ɖʱ
ण	ਣ	ण	ɳ
त	ਤ	त	t̪
थ	ਥ	थ	t̪ʰ
द	ਦ	द	d̪
ध	ਧ	ध	d̪ʱ
न	ਨ	न	n
प	ਪ	प	p
फ	ਫ	फ	pʰ
ब	ਬ	ब	b
भ	ਭ	भ	bʱ
म	ਮ	म	m
य	ਯ	य	j
र	ਰ	र	ɾ
ल	ਲ	ल	l
व	ਵ	व	ʋ/ w
श	ਸ਼	श	ʃ
ष		ष	ʂ / ʃ
स	ਸ	स	s
ह	ਹ	ह	ɦ
ड़	ੜ	ड़	ɽ̣
ढ़		ढ़	ɽʰə
	ਲ਼		!

Table 1(b): Consonants with their transcription symbols for Hindi, Punjabi and Nepali

2. Text data collection and corpus design

The corpora used in this study were selected by a method that makes it a reasonable representative of a given language. The text materials for these languages were collected from three domains i.e. Defense, General information and News items. This process involved the collection of words for various languages by processing the online news, defence websites and general purpose information blogs, stories etc. In this way for statistical analysis, we have created the large text corpus of approximately 500000 words for each Hindi ,Punjabi, Indian English and Nepali language.

3. Statistical analysis and Observation

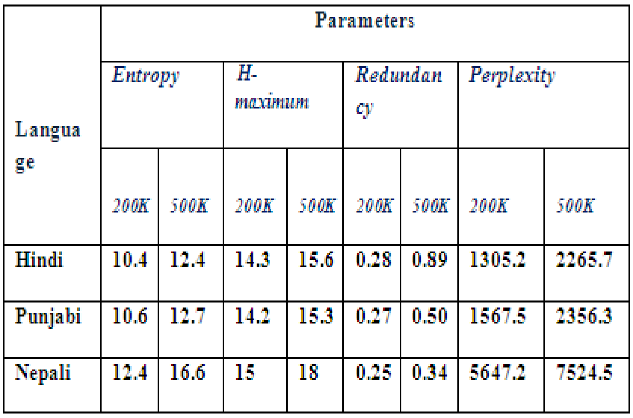
In the present study, the statistical analysis have been extended on a bigger database of 500000 words as compared our earlier study on a corpora of 200000 words.

(A) Entropy and Perplexity

In general, entropy is used to measure uncertainty associated with a random variable. In NLP, speech recognition and computational linguistics, entropy is used as the measure of information [2]. It finds application in various fields like how much information is there in a particular grammar, how well given grammar matches a given language etc. The information is more predictable if entropy is low whereas less predictable for higher entropy. The formula used for computation of entropy is as follows:

$$\mathbf{H\ (X) = \sum_{x \in X} -p(x)\ log_2\ p(x)}$$

where X is random variable ranging from 1 to number of word types in the language in this case. Similarly, relative entropy, maximum entropy, redundancy and perplexity have been also calculated for all these three languages. H-Maximum or maximum entropy is obtained when the probabilities of all words in the corpus are same. The percentage of redundancy can be used to analyze whether the data can be compressed and stored in less number of bits or not comparatively. Whereas the perplexity is the measurement in information theory. In NLP, perplexity is a common way of evaluating language models. It measures the goodness of model[3] . Lower perplexity means the corpora is more predictable. Table 2 represents the values obtained for the above defined parameters for three languages from the corpus of 500000 words of each language:



Langua ge	Parameters							
	Entropy		*H-maximum*		*Redundancy*		*Perplexity*	
	200K	*500K*	*200K*	*500K*	*200K*	*500K*	*200K*	*500K*
Hindi	10.4	12.4	14.3	15.6	0.28	0.89	1305.2	2265.7
Punjabi	10.6	12.7	14.2	15.3	0.27	0.50	1567.5	2356.3
Nepali	12.4	16.6	15	18	0.25	0.34	5647.2	7524.5

Table 2: Table showing entropy, H-maximum, Redundancy and Perplexity of three languages computed from corpus of 200k and 500k words

The above statistical data helps to find the number of bits required to encode the word in a given language. From table 2 it can be observed that Hindi has high redundancy values compared to other languages and Nepali has the lowest redundancy value therefore the efficiency of Hindi is lowest among all languages and Nepali has highest efficiency . It has also been observed that Nepali has higher entropy whereas other languages have almost same entropy. Higher entropy means the given information is more uncertain or harder to predict. Also, lower the entropy higher the data compression rate therefore data compression is expensive for Hindi and Punjabi in comparison with Nepali. Nepali has highest value of perplexity among all.

(B) Vocabulary Growth Rate

In figure 1 , Function V (N) is the number of word tokens extracted from the total data base N, collected for Punjabi, Nepali and Hindi. For observed values of the different language vocabulary growth rate has been drawn which has been shown in graph by HI(O),PU(O) and NE(O). The curve fitting method is used to obtain the growth curve which gives expected values of vocabulary size , interpolation is done to get the expected value E[V(N)] for smaller sample sizes of N and extrapolation to get the expected value E[V(N)] for larger sample sizes of N. It has been observed that the curve may provide excellent fit for Hindi and Punjabi languages but in Nepali language due to variable frequency distributions of words, there is a variation in the expected and observed values.

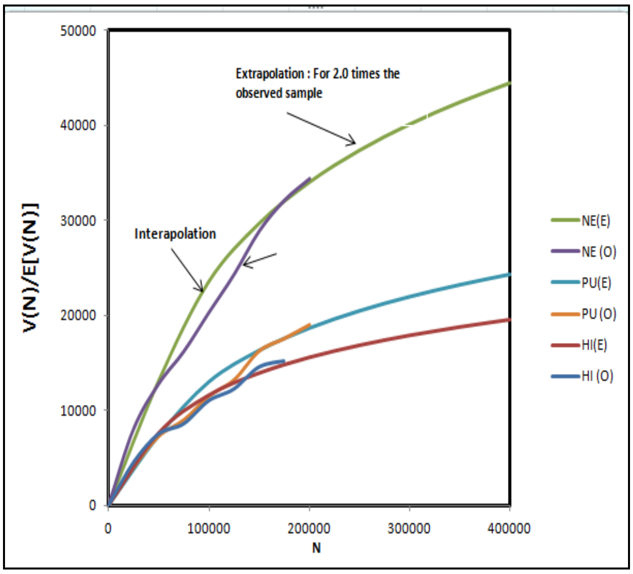


Figure 1: Vocabulary growth rate for Hindi, Punjabi and Nepali

Vocabulary graph gives the growth curve for extrapolating to 2 times the observed sample size. It has been observed that the curve may provide excellent fit for Hindi language but due to variable frequency distributions of words in Punjabi and Nepali, there is a variation in the expected and observed values [4,5,6].

3. Development of Language Model

In this experiment models for each language i.e. Punjabi, Nepali and Indian English have been developed. Comparison among all the three models has also been done based on the various monograms, bigrams and trigrams with their corresponding probabilities. The implementation algorithm has been developed, for training the database for finding various monograms, bigrams and trigrams present in the database with their corresponding probabilities using the formulas for calculation of n-gram probabilities [8]. An algorithm has also been developed for computation of probability of the text which can be taken from any source such as website or a written script etc. Usage of JAVA platform has provided a special effect for dealing with a very large database with high computational speed. Furthermore various enhancive concepts like Smoothing, Discounting, Backoff, and Interpolation have been included for designing of the model. The language model is based on the conditional probability and this is because, we compute the probability of the occurrence of any word when a sequence of words has already occurred. The probabilistic language model can be illustrated as $P\ (W_1 \dots W_m) = \Pi\ P\ (W_i/W_1 \dots .W_{i-1})$, for i=1 to m, where W_i is the ith word in the sentence/utterance, Π defines the multiplication function and P represents the conditional probability, which gives the general equation for probability calculation of any sentence. [8]

In trigram model, full stop (.), question mark (?) and exclamatory sign (!) symbols have been replaced by <s> <s> so that the new sentence should not affect the ending word of the previous sentence and vice versa. The Trigram model has been trained for the three non parallel databases of these languages by using following algorithm:

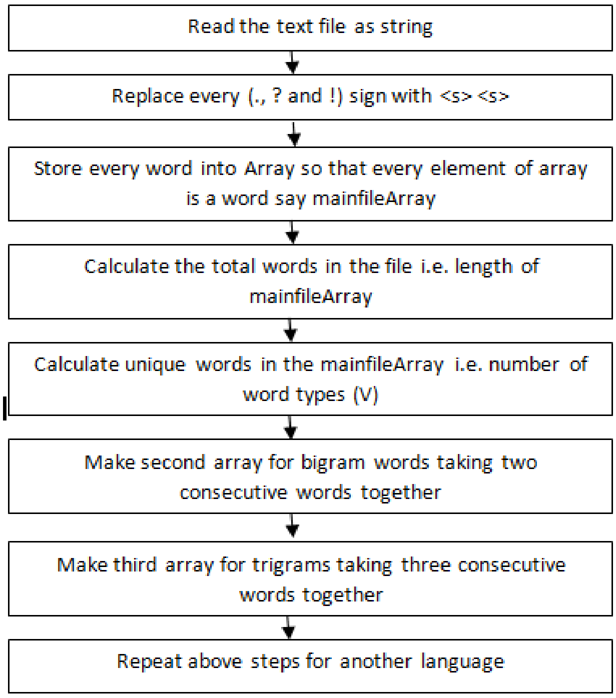
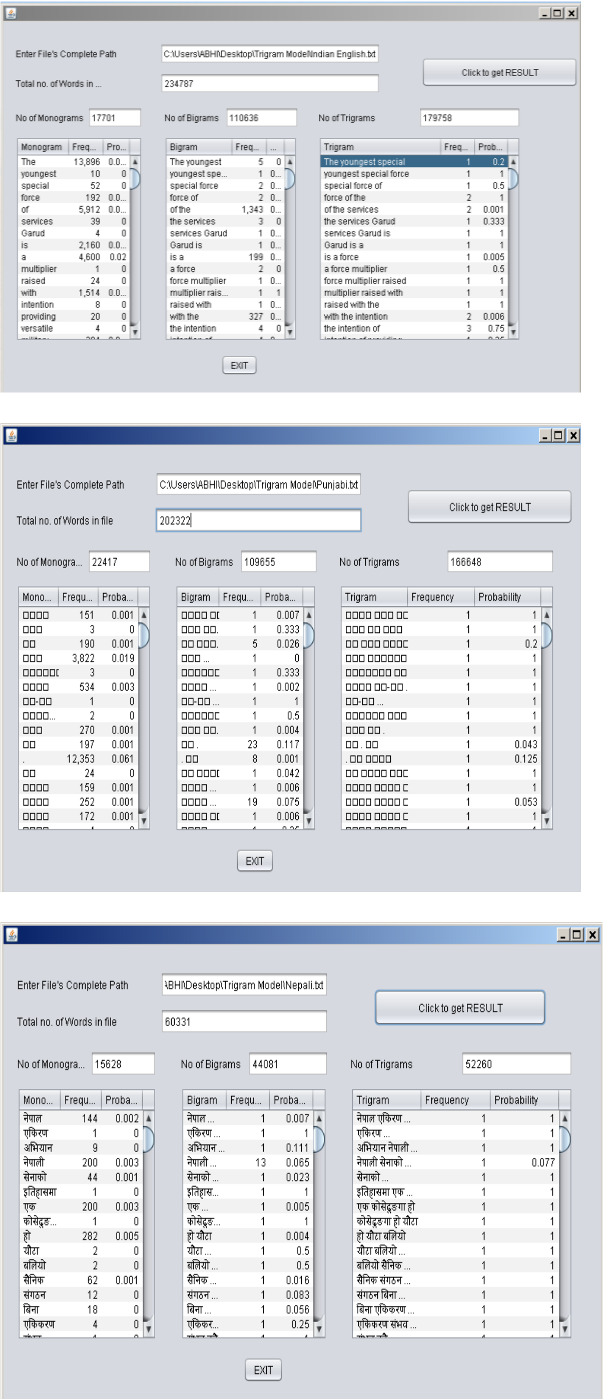

Read the text file as string
Replace every (., ? and !) sign with <s> <s>
Store every word into Array so that every element of array is a word say mainfileArray
Calculate the total words in the file i.e. length of mainfileArray
Calculate unique words in the mainfileArray i.e. number of word types (V)
Make second array for bigram words taking two consecutive words together
Make third array for trigrams taking three consecutive words together
Repeat above steps for another language

Figure 2: Flowchart of algorithm used for training of Trigram Model

For non-parallel databases, the model has been tested for Indian English, Nepali and Punjabi languages. The results have been gathered for getting the probability of the sentence/text. GUI has been created for both the models (training and testing) and tested for all the three languages viz. Indian English, Nepali and Punjabi(as shown in figure3). The GUI for Punjabi language show undefined symbols as no Interface supporting the Punjabi language has been found, though the calculation and computational results have been found to be correct.



Enter File's Complete Path: C:\Users\ABHI\Desktop\Trigram Model\Indian English.txt

Total no. of Words in ...: 234787

No of Monograms: 17701 | No of Bigrams: 110636 | No of Trigrams: 179758

Monogram	Freq...	Pro...
The	13,896	0.0...
youngest	10	0
special	52	0
force	192	0.0...
of	5,912	0.0...
services	39	0
Garud	4	0
is	2,160	0.0...
a	4,600	0.02
multiplier	1	0
raised	24	0
with	1,514	0.0...
intention	8	0
providing	20	0
versatile	4	0

Bigram	Freq...	...
The youngest	5	0
youngest spe...	1	0...
special force	2	0...
force of	2	0...
of the	1,343	0...
the services	3	0
services Garud	1	0...
Garud is	1	0...
is a	199	0...
a force	2	0
force multiplier	1	0...
multiplier rais...	1	1
raised with	1	0...
with the	327	0...
the intention	4	0

Trigram	Freq...	Prob...
The youngest special	1	0.2
youngest special force	1	1
special force of	1	0.5
force of the	2	1
of the services	2	0.001
the services Garud	1	0.333
services Garud is	1	1
Garud is a	1	1
is a force	1	0.005
a force multiplier	1	0.5
force multiplier raised	1	1
multiplier raised with	1	1
raised with the	1	1
with the intention	2	0.006
the intention of	3	0.75

Enter File's Complete Path: C:\Users\ABHI\Desktop\Trigram Model\Punjabi.txt

Total no. of Words in file: 202322

No of Monogra...: 22417 | No of Bigrams: 109655 | No of Trigrams: 166648

Mono...	Frequ...	Proba...
□□□□	151	0.001
□□□	3	0
□□	190	0.001
□□□	3,822	0.019
□□□□□□	3	0
□□□□	534	0.003
□□-□□	1	0
□□□□...	2	0
□□□	270	0.001
□□	197	0.001
.	12,353	0.061
□□	24	0
□□□□	159	0.001
□□□□	252	0.001
□□□□	172	0.001

Bigram	Frequ...	Proba...
□□□□ □□	1	0.007
□□□ □□.	1	0.333
□□ □□□.	5	0.026
□□□ ...	1	0
□□□□□□	1	0.333
□□□□ ...	1	0.002
□□-□□ ...	1	1
□□□□□□	1	0.5
□□□ □□.	1	0.004
□□ .	23	0.117
. □□	8	0.001
□□ □□□□	1	0.042
□□□□ ...	1	0.006
□□□□ ...	19	0.075
□□□□ □□	1	0.006

Trigram	Frequency	Probability
□□□□ □□□ □□	1	1
□□□ □□ □□□	1	1
□□ □□□ □□□□	1	0.2
□□□ □□□□□□	1	1
□□□□□□□ □□	1	1
□□□□ □□-□□ .	1	1
□□-□□ ...	1	1
□□□□□□ □□□	1	1
□□□ □□ .	1	1
□□ . □□	1	0.043
. □□ □□□□	1	0.125
□□ □□□□ □□□	1	1
□□□□ □□□□ □	1	1
□□□□ □□□□ □	1	0.053
□□□□ □□□□ □	1	1

Enter File's Complete Path: \BHI\Desktop\Trigram Model\Nepali.txt

Total no. of Words in file: 60331

No of Monogra...: 15628 | No of Bigrams: 44081 | No of Trigrams: 52260

Mono...	Frequ...	Proba...
नेपाल	144	0.002
एकिरण	1	0
अभियान	9	0
नेपाली	200	0.003
सेनाको	44	0.001
इतिहासमा	1	0
एक	200	0.003
कोसेढुङ...	1	0
हो	282	0.005
यौटा	2	0
बलियो	2	0
सैनिक	62	0.001
संगठन	12	0
बिना	18	0
एकिकरण	4	0

Bigram	Frequ...	Proba...
नेपाल ...	1	0.007
एकिरण ...	1	1
अभियान ...	1	0.111
नेपाली ...	13	0.065
सेनाको ...	1	0.023
इतिहास...	1	1
एक ...	1	0.005
कोसेढुङ...	1	1
हो यौटा	1	0.004
यौटा ...	1	0.5
बलियो ...	1	0.5
सैनिक ...	1	0.016
संगठन ...	1	0.083
बिना ...	1	0.056
एकिकर...	1	0.25

Trigram	Frequency	Probability
नेपाल एकिरण ...	1	1
एकिरण ...	1	1
अभियान नेपाली ...	1	1
नेपाली सेनाको ...	1	0.077
सेनाको ...	1	1
इतिहासमा एक ...	1	1
एक कोसेढुङगा हो	1	1
कोसेढुङगा हो यौटा	1	1
हो यौटा बलियो	1	1
यौटा बलियो ...	1	1
बलियो सैनिक ...	1	1
सैनिक संगठन ...	1	1
संगठन बिना ...	1	1
बिना एकिकरण ...	1	1
एकिकरण संभव ...	1	1

Figure 3: Screenshots of total Monograms, Bigrams and Trigrams for Indian English, Punjabi and Nepali

The above screen shots show the results for First GUI created which finds the total no. of words in the database, total Monograms, Bigrams & Trigrams and their counts. The GUI has been tested for all the three languages viz. Indian English, Nepali & Punjabi. The GUI for Punjabi language show undefined symbols as no Interface supporting the Punjabi language found. Though the calculation and computational results have been found correct.

The comparative analysis of the monogram, bigram and the trigram model designed for all the three languages has been shown in the following table:

Sr. No.	Language	Total words in Database	No. of Monograms	No. of Bigrams	No. of Trigrams
1	Indian English	**200000**	**17701**	**110636**	**179758**
2	Nepali	**200000**	**15628**	**104081**	**152260**
3	Punjabi	**200000**	**22417**	**109655**	**166648**

Table 3: Number of Monograms, Bigrams and Trigrams available in the database of all the three languages

4. Implementation of Trigram Model for Language Identification of Parallel Database on other Languages.

The model was tested on the parallel corpus (text database) of French and English using following steps (shown in figure4).

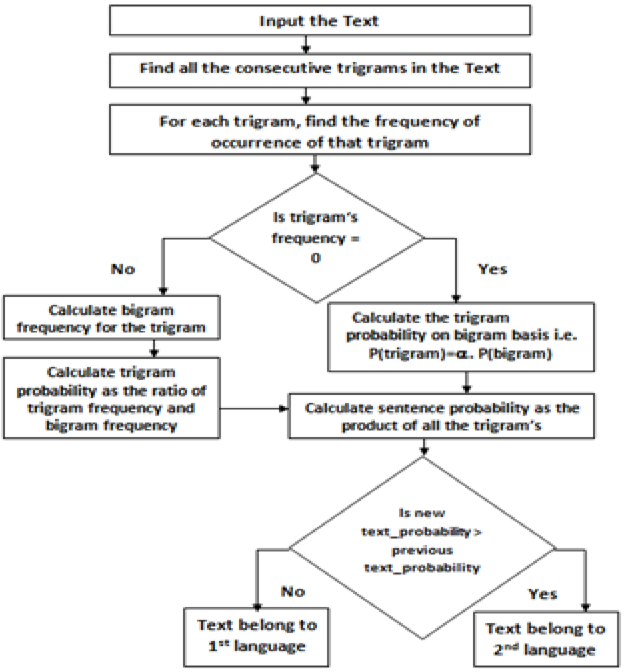

Input the Text
Find all the consecutive trigrams in the Text
For each trigram, find the frequency of occurrence of that trigram
Is trigram's frequency = 0
No
Yes
Calculate bigram frequency for the trigram
Calculate trigram probability as the ratio of trigram frequency and bigram frequency
Calculate the trigram probability on bigram basis i.e. P(trigram)=α. P(bigram)
Calculate sentence probability as the product of all the trigram's
Is new text_probability > previous text_probability
No
Yes
Text belong to 1st language
Text belong to 2nd language

Figure 4: Block diagram of steps used for language identification

It was found very effective in distinguishing the language. French database consisted of 34628 word tokens and English database of 38198 word tokens. The model was implemented to identify the languages based on same sample sentences and found to be quite successful. The probabilities were quite different for example a sentence in English was 7.070E-19 as compared to 1.9065E-19 for French.[7]

5. Conclusion

In this paper we have described a variety of statistical analyses of large text corpora of Hindi, Punjabi and Nepali. The paper shows that Nepali language is more complex among these languages. The data presented above can be used for the development and improvement of statistical procedures of linguistic analysis and will make possible the construction of more satisfactory mathematical models of language. .We have also proposed a simple yet intuitive trigram model using JAVA platform which can handle very large database and gives a greater efficiency as compared to other languages like C, C++, and Python etc.; for calculating the probability of text in the database and thus identifying the language whether the given text belongs to English language or French Language. The language model so far designed is being used for the purpose of language identification and web content analysis. The model will also be used for higher level recognition of spoken languages and speaker recognition along with their linguistic background.

6. Acknowledgement

The authors would like to acknowledge the suggestions and help received from Mr. A.K. Dhamija and Mr. Gautam Kumar in recording and analysis of the speech samples collected from speakers of different linguistic backgrounds. Our special thanks to Dr. Harsh Vardhan Kamrah and Mrs. Neelima Kamrah for their support and encouragement. We also thank SAG, DRDO for providing partial financial support for conducting these studies.

REFERENCES

1. http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India
2. G.S Lehal, Renu Dhir, Ritu Lehal, “Corpus based statistical analysis of printed Punjabi text” in proceedings of International Conference on Knowledge Based Computer Systems, 17-19 Dec 1998, NCST, Mumbai.
3. Akshar Bharthi, Rajeev Sangal and Sushma M Bendre, “Some Observations Regarding Corpora of Indian Languages” Proceedings of KBCS-98, 17-19 Dec 1998, Mumbai.
4. Sunita Arora, Karunesh Arora, S.S Agrawal, “Statistical Analysis of Hindi BTEC Speech Database” OCOCOSDA ,9-12 December 2012, Macau, China
5. Sunita Arora, Karunesh Arora, Aman Suneja, S.S Agrawal, “Statistical Analysis of Text and Speech Sounds from Parallel Corpus of Indian Languages” OCOCOSDA, Singapore, 2003.
6. Shweta Bansal, Minakshi Mahajan , S.S. Agrawal, “Determination of Linguistic Differences and Statistical Analysis of Large Corpora of Indian Languages” OCOCOSDA ,Nov. 2013, Gurgaon , India
7. Abhimanue Mandal, S.S Agrawal “Development of Preliminary model for language Identification purpose” KIIT Research Journal ,Vol.3 March 2013
8. Vatanen, Tommi and Väyrynen, Jaakko J. and Virpioja, Sami. “Language Identification of Short Text Segments with N-gram Models.” European Language Resources Association
9. Yew Choong Chew, Yoshiki Mikami, Robin Lee. “ Language Identification of Web Pages Based on Improved N-gram Algorithm.” IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011