Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.54 – 60

RESEARCH ARTICLE

A Detailed Study on Indian Languages Text Mining

M. Hanumanthappa[1], M. Narayana Swamy[2]

[1]Department of Computer Science, Bangalore University, Bangalore Karnataka, India
[2]Department of Computer Science, Presidency College, Bangalore, India
[1] hanu6572@hotmail.com, [2] narayan1973.mns@gmail.com

Abstract— India is a country with huge population of over hundred and twenty seven core, who speaks different languages. Only 5% of Indian population can effectively communicate in English and rest 95% are comfortable with their regional languages. India is certainly one of the multilingual nations in the world today.

In the Constitution of India, a provision is made for each of the Indian states to choose their own official language for communicating at the state level for official purpose. To penetrate the benefits of Communication and Information Technology up to common masses, the content is available in Indian language. In India, we are starting to see a growth in consumption of Indian language content, because of growth of electronic devices and technology. As these devices get cheaper, the internet is accessible in the smaller towns and rural parts of the country. So because of growth of internet, the demand for content in Indian language is also been rising. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. Not much work has been done in Indian languages text processing.

The objective of this paper is to understand the following: Growth of data in Indian languages, Need of text mining for Indian languages, Literature survey on Indian language text mining, Application and so on

Keywords— TDIL, W3C, LLP, LIP, CLIP, Zipf's Law

I. INTRODUCTION

India has more languages than any other country in the world. So India is a multi-linguistic, multi-script country with 23 official languages and 11 written script forms. About a billion people in India use these languages as their first language. English, the most common technical language, in the government, and the court system, but is not widely understood beyond the middle class and those who can afford formal, English-language education.

People throughout the world have been using computers and Internet in their own languages. Even though the India is multi- linguistic country, But the Indian users are compelled to use them in English. In India among men 72 per cent do not speak English, 28 per cent speak at

least some English, and 5 percent are fluent. Among women, the corresponding proportions were 83 per cent, 17 per cent, and 3 per cent [1]. But the Indian engineers and scientists, dominant force in the IT world have faced criticism for neglecting the needs of common man from their own region. To fill this gap many organizations like Microsoft, TDIL are trying to introduce the software in regional languages to take IT solutions to the rural areas. To bridge the digital divide and leverage the power of information and communication technology it is essential to reach out to masses and overcome the language barrier.
The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. So in future Indian language text processing is required.

II. GROWTH OF DATA IN INDIAN LANGUAGES

Languages cannot be preserved by making dictionaries or grammars. Languages live if people who speak the languages continue to live. So we need to look after the well being of the people who use those languages. So the scientists and the government of India are working towards this for planning of development where language is taken as one factor.
The first step in this direction was the launch of TDIL (Technology Development for Indian Languages) Programme in 1991 by Ministry of Information Technology to develop information processing tools to facilitate human machine interaction in Indian Languages and to create and access multilingual knowledge resources and integrating them to develop innovative products and services for users. The next milestone has been the setting up of Resource Centres for Indian Language Technology Solutions. These centers will develop technologies for providing solutions with citizen interface in Indian languages selectively and thus covering all Indian languages.

World Wide Web consortium (W3C) Promoting the Multi-language usage for Mobile and Internet: W3C is a Government of India initiative with the Department of Information Technology. As part of its efforts to ensure that core Web standards meet global needs, this dept has been created to promote W3C standards all over India in all twenty-two constitutionally recognized Indian languages. [2]

Some other projects for Indian languages are as follows:
1. Project Bhasha is a comprehensive program which aims to localize Microsoft's flagship products, Windows and Office in 12 Indian Languages
2. Bhasha Online Community portal is India's leading community for Indian language computing.
3. The Indic Language Input tool is a phonetic based keyboard which facilitates users to input localized text easily and quickly. The Indic Language Input tool is available in two versions; the desktop version enables the user to enter Indian language text directly into any application running on Windows, such as Microsoft Word or Outlook. The web version allows the user to enter text on any web page - such as live mail or Windows Live Messenger - without requiring software download. The tool currently supports ten languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu.
4. The desktop computer software is localized for the major world languages. The Microsoft Local Language Program (LLP) is part of worldwide initiative dedicated to providing desktop computer software in their native language. Beyond providing fully localized versions of Microsoft windows and Microsoft Office in nearly 40 languages, Microsoft currently supports 95 languages through the LLP. The Local Interface Packs (LIPS) and Caption Language Interface Packs (CLIPS), which are part of the Microsoft Local Language Program. Today Microsoft Windows and Microsoft Office, reaching more than 90

percent of the global population. In addition to providing these Local Language Packs, Microsoft and Google also provide online dictionaries, translation tools and localized versions of our developer tools. Ultimately these tools and resources help support language preservation and translation, which can lead to better economic opportunities through giving more people access to technology in their own language.

5. Language Interface Packs (LIPs) in 12 Indian languages- Assamese, Bengali, Gujarati, Hindi, Kannada, Konkani, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu- for MS Office and Windows. A total of 45 additional soft (virtual) keyboards, which are free to download, are also available in these 12 languages.

In 2009 and 2010, there was an industry attempt to evolve a standard where several companies joined to work with the Centre for Excellence in Wireless and Internet. In June 2013, DeitY had also created a repository of fonts for all 22 constitutionally recognized languages through TDIL. In July 2013, Technology Development for Indian Languages Programme (TDIL) of DeitY (Department of Electronics and IT) had developed Urdu language fonts and keyboard drivers for Windows and Android [3].This is a big change, when Internet access was still a largely urban, English-first phenomenon. However, while sites like Google launched a Hindi homepage in 2009, and now support Gujarati, Tamil, Marathi and Bengali; and others like Twitter started to support Hindi in 2011, there are still relatively few Indian language-only sites among the top 100 sites being visited from India.

One of the older and better known Indian language search engines is Raftaar.in, which has content in Hindi. Indiblogger aggregates links to Indian blogging sites, and you can choose between Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu content on the site. The site lists nearly 2,000 Hindi blogs, while the other languages account for nearly another 2,500, with Tamil leading with 975 blogs. Popular sites in Hindi include Facebook, YouTube, Wikipedia, Twitter, Google, Wordpress, Bing, Blogger, Vube and Indiatimes, which shows that the audience wants the same content as everybody else, and not just local language websites. This means that there is a huge, untapped audience which wants to read the predominantly English language websites in their native language. One of the few e-commerce sites that has already adopted Hindi is snapdeal, which launched Hindi and Tamil versions, and plans to roll out more versions.

The diversity in languages across the Indian landscape is one of the key challenges in taking Internet to the masses. While the state governments have been working on several e-governance platforms for the citizens, making information available in regional languages to enable efficient use of these services is as much a last-mile challenge as Internet infrastructure in India.

III.NEED FOR TEXT MINING

Impact of Information Technology was felt as early in 1970s. Solutions towards adaptation of rapidly growing Information Technology for Indian languages were developed. Input-output problems and coding schemes were analysed. In 1990-91, Government launched the program on TDIL (Technology Development of Indian Languages) under which projects were supported for development of corpora, OCR, Text-to-Speech, machine translation and generic software for Information processing. Standards for keyboard layout and internal Code for Information Interchange were also evolved. This resulted into confidence in having solutions for Information processing in Indian languages[4].

So the availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. The web content in Indian language has been increasing .There have been many portals, which host large amounts of the Indian language content. However, we argue that the data is being underutilized due to the unavailability of Indian language text mining methods. Therefore text mining is required for Indian regional languages.

IV.LITERATURE SURVEY

From the literature survey we have noticed that Not much work has been done in Indian languages text processing. Here we made an attempt to summarize the research work on Indian languages

[5] In this paper, they have reported work on Name Entity Recognition (NER) model , keyword extraction and topic tracking for Punjabi language. They have prepared the system with the combination of two approaches, i.e. combining a number of features from NER and keyword extraction to generate effective topic tracking system. The language dependent features and language independent features are formed and analyzed. The approach uses the gazetteer lists created from the dictionary with part of speech tagging, morphological analyzer, Punjabi vocabulary. Hence, NEs such as date/ time, location, person name, organization, designation and keywords from title, cue phrase and high frequency noun are extracted. The system shows good results for all features independently and the total results for the system are improved with the combination of these features resulting in effective topic tracking system.

[6]Part of speech tagging plays a vital role in natural language processing. This paper presents a reasonably accurate POS tagger for Kannada language. Part of Speech tagging helps in the creation process of a parser.

[7]In this paper, they have used Domain Based Ontology for the Classification of Punjabi Text Documents..This is the proposed algorithm for the classification of Punjabi Text Documents. in this approach they have made an initiation to create an ontology for Punjabi Language by creating Sports Based Ontology in Punjabi that consist of class related terms. One of the major advantages of this ontology based classification is that it does not need Training Data i.e. Labeled Documents to classify the documents, whereas other Classification Techniques such as KNN technique, Naive Bayes Algorithm, Association Based Classification etc. need Training Set or Labeled Documents to train the classifier to do the classification of the unlabelled documents.

[8]They have worked on single- document opinion summarization for Bengali. The novelty of the proposed technique is the topic based document-level theme relational graphical representation. This is the first attempt on opinion summarization for Bengali. The approach presented here is unique in every aspect as in literature and for a new language like Bengali.

[9] Most of the lexical resources used in pre processing and processing such as Punjabi stemmer, Punjabi nouns normalizer, Punjabi proper names list, common English-Punjabi nouns list, Punjabi stop words list, Punjabi suffix and prefix list etc. had to be developed from scratch as no work was done previously in that direction. For developing these resources an in-depth analysis of Punjabi corpus, Punjabi dictionary and Punjabi morph had to be carried out using manual and automatic tools. This is first time that these resources have been developed for Punjabi and these can be beneficial for developing other Natural language processing applications for Punjabi.

[10] They have worked on pre categorized data; there are good classifiers which can provide the necessary classification. It can be extended to first classify a given document and then create a summary. There is no standard stop word list for Kannada, or methods to do that.

Hence a given procedure in this work can be used as a stop word removal method. The summarizer can be used as a tool in various organizations such as Kannada Development Authority, Kannada Sahitya Parishath etc.

[11] They have presented the design and working of tourist decision assistance system that helps tourists in selecting places to visit based on their preference including locations. The system performs with an accuracy of 94 percent on unseen dataset. A major problem encountered by them was the presence of missing values in many documents. We also had to solve the problem of handling a mix of Unicode compatible and non-Unicode compatible source documents. The documents digitized earlier were encoded using ISCII (Indian Script Code for Information Interchange) fonts which are not Unicode compatible.

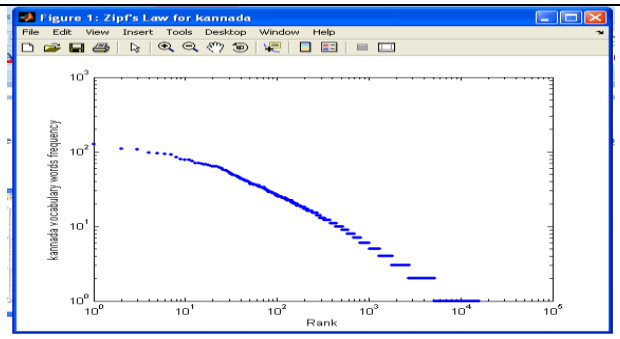
[12] They have conducted on Telugu documents using Naive Bayes classifier. They have a base system on which a variety of further explorations can be carried out, both from the linguistic point of view and statistical point of view. With the increasing availability of large scale data, affordable memory and computing power, deeper analysis in both linguistic and statistical sense are becoming possible. Morphological analysis and stemming would be high on the agenda. Role of Phrases and Collocations would be worth exploring. Impact of Syntactic Parsing and Word Sense Disambiguation may be explored. Stop word removal and other usual clean up techniques can be incorporated.

Table : 1

Language	Kannada	Tamil	Telugu
Documents	100	100	100
Tokens	26315	20360	18427
Vocabulary	20417	15941	14652

V. FUNDAMENTAL PROPERTY OF LANGUAGES

The most fundamental property of languages is the one known as Zipf's law. For any language, if we plot the frequency of words versus their rank for a sufficiently large collection of textual data, we will see a clear trend, which resembles a power law distribution.

	Zipf's law
Kannada	

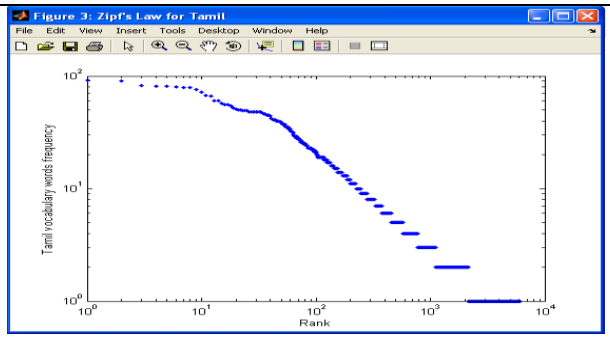
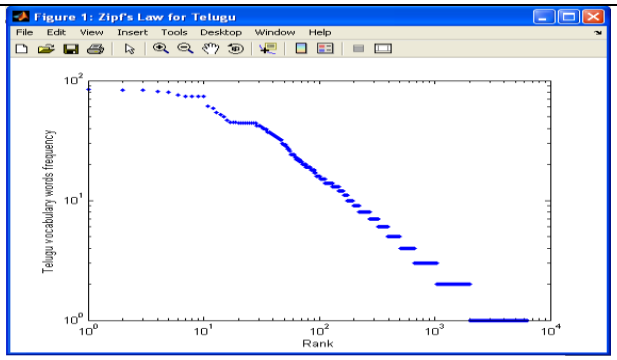
Tamil	
Telugu	

Table 2

Our experiment on Kannada, Tamil and Telugu corpus statistics is illustrated by Zipf's law[14].The properties of large volume of text are generally referred to as corpus statistics. This data collection comprises 300 documents. The basic statistics of corpus is shown in the table 1.

As seen from the table 2, in the low rank extreme of the curve, which are clearly separated from the rest of the words, These are the most frequently used words in our considered data collection.

VI. APPLICATIONS

According to a latest study conducted by Internet and Mobile Association of India and IMRB International, regional content availability can boost the growth of Internet in India by 24%. The study said that in 2013 the regional language content users grew by 15% to 71.8 million from 45 million in 2012[13].

It is well known that one result of the Internet's rapid growth has been a huge increase in the amount of unstructured data in regional language. Text mining application uses unstructured textual information. Some examples of practical applications of text mining techniques include

- Spam filtering
- Creating suggestion and recommendations (like amazon)
- Monitoring public opinions (for example in blogs or review sites)
- Customer service, email support
- Automatic labeling of documents in business libraries
- Measuring customer preferences by analyzing qualitative interviews
- Fraud detection by investigating notification of claims
- Fighting cyberbullying or cybercrime in IM and IRC chat

VII. CONCLUSIONS

Not much work has been carried out on Indian language texts. Summary: India's growing focus on Internet services being provided in regional languages. So the availability of textual data of various Indian regional languages in electronic form has accelerated. So in future Indian language text processing is required.

The text document may contain a few structured fields, and also unstructured text components, without knowing what could be in the documents, it is difficult to formulate effective queries for analysing and extracting useful information from the data. To compare the documents and rank the importance and relevance of the document the users need tools. Therefore, text mining has become popular and essential for Indian languages

REFERENCES

[1] " Human Development In India" . OUP. 2005.
[2] "ANNUAL REPORT 2010-11", IAMAI
[3] A Big Need for Indic-Language Solutions
[4] Developing Information Technology Solutions in Indian Languages: Pros and Cons
[5] "Topic Tracking For Punjabi Language", Kamaldeep Kaur and Vishal Gupta, Computer Science & Engineering: An International Journal (CSEIJ), Vol.1, No.3, August 2011 DOI : 10.5121/cseij.2011.1304 37
[6] " POS Tagger for Kannada Sentence Translation", Mallamma V Reddy and Dr. M. Hanumanthappa, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 1, May-June 2012 ISSN 2278-6856
[7] "Algorithm for Punjabi Text Classification", Nidhi and Vishal Gupta, International Journal of Computer Applications (0975 – 8887) Volume 37– No.11, January 2012
[8] "Topic-Based Bengali Opinion Summarization ", Amitava Das and Sivaji Bandyopadhyay, Coling 2010: Poster Volume, pages 232–240, Beijing, August 2010
[9] "Automatic Punjabi Text Extractive Summarization System", Vishal Gupta and Gurpreet Singh Leha, Proceedings of COLING 2012: Demonstration Papers, pages 191–198, COLING 2012, Mumbai, December 2012.
[10] "Document Summarization In Kannada Using Keyword Extraction" Jayashree.R, Srikanta Murthy.K and Sunny.K, Computer Science & Information Technology (CS & IT)
[11] "Oriya Language Text Mining Using C5.0" Algorithm Sohag Sundar Nanda, Soumya Mishra, Sanghamitra Mohanty ISSN : 0975-9646
[12] "Automatic Categorization of Telugu News Articles", Kavi Narayana Murthy
[13 Press Coverage Mobile and social media drive regional language content Mon, 09 Jun 2014 - Business Standard
[14] Narayana Swamy and, Hanumanthappa "Indian Language Text Representation and Categorization using Supervised Learning Algorithm" ICICA '14 Proceedings of the 2014 International Conference on Intelligent Computing Applications, IEEE Computer Society Washington, DC, USA , 2014 ISBN: 978-1-4799-3966-4

