International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 5, Issue 3 (May-June 2017), PP. 122-124

STUDY OF MARATHI TEXT INPUT MECHANISM FOR SMART PHONES AND MARATHI WORD PREDICTION

[1] Yashoda Bhat, [2] Prajakta Akerkar, [3] Mona Deshmukh
[1,2]Student, [3] Professor
[1,2] Dept. of MCA, VESIT, Mumbai, Maharashtra, India
[3] Dept. of MCA, VESIT, Mumbai, Maharashtra, India
[1]yashoda.bhat@ves.ac.in

Abstract— **In this paper we have studied the Corpus of Marathi Word Frequencies from Touch-Screen Devices Using Swarachakra Android Keyboard and the online a corpus containing word frequencies of Marathi texts that were actually typed by 27,474 users using the Android version of the Swarachakra Marathi keyboard on their mobile devices between August 2013 and September 2014 and a new mechanism for the word autocorrection and autocompletion is also suggested . We hope and expect that this will be useful for future researchers, particularly those involved in word completion and auto-correction of user errors.**

Keywords—**Swarachakra,auto-correction,word completion,n-gram algorithm.**

I. INTRODUCTION

letters that are possible to complete those wors. Each key press results in a indicator rather than again sequencing through the same group of "letters" it represents, in the same, invariable order. Predictive input may allow for an whole word to be input by single key press. Predictive text makes adequate use of fewer device keys to input writing into a text message, an e-mail, an address book, a calendar, and the like[1].

Text input researchers have proposed different methods for creating corpora artificially by picking data from several sources .

Different types of corpora datasets are available for many languages including Marathi [2,3,4]. However, many of these corpora are have been derived from books, newspapers, or online sources. One could argue that these are “formal corpora” as the texts have been written by expert writers (book authors, journalists etc.) and edited by professional editors before publication. Moreover, these corpora have been primarily typed on desktop computers, and in case of Marathi, usually by professional typists (as typing in Marathi by users who are not professionally trained typists is relatively rare). What common people type using a mobile device though, may be completely different. Firstly, it may be influenced by several factors such as the devices used for writing,

II. LITERATURE STUDY

EMILLE [1] and FIRE [3] are the two largest and most widely used corpora in Marathi. The EMILLE corpus was constructed collaboratively by the EMILLE project (Enabling Minority Language Engineering), Lancaster University, UK, and Central Institute of Indian Languages (CIIL), Mysore, India. It has monolingual (14 languages – 92,799,000 words), parallel (200,000 words of parallel text for 5 languages) and annotated (2 languages) corpora. The corpus was not balanced for genres or data types, aim being to collect all electronic data available. The corpus includes scientific writing, fiction, educational texts, and news data from the web. The spoken data was gathered mainly from BBC Asian Network. Under the Forum for Information Retrieval Evaluation (FIRE) workshop, collections of data were made available in 7 languages – Bengali, English, Gujarati, Hindi, Marathi, Tamil and Telugu. FIRE is part of the effort by the Information Retrieval Society of India (IRSI) to promote research in the area of Information Retrieval. The Marathi collection mainly contains newspaper data from the popular Marathi dailies – Maharashtra Times, Sakal and Loksatta. Wikipedia is the third major corpus for Marathi that is freely available for download [8]. This corpus contains text and titles of the wiki pages in Marathi Wikipedia along with associated metadata

III. IMPLEMENTATION

Swarachakra [5] is a logically structured, open-source keyboard for text input in Indic scripts for touchscreen devices, which has been developed by several contributors. The layout of the keyboard is based on the structure of the Indic scripts. A unique aspect of Swarachakra is that when the user touches a consonant, it throws up a circle (chakra) that shows a preview of the most popular vowel modifiers combined with that consonant (the consonant ध has been touched and the chakra shows a preview of धा, धध, धी... etc.). The user needs to slide the finger in the direction of the desired consonant + vowel modifier combination.

Fig.1.Example of swarachakra keyboard

This effectively converts what would take two or more taps on a typical keyboard into a single slide .Because the keyboard is logically ordered, this further reduces the cognitive load that users often face while typing in Indic scripts. As of now, Swarachakra does not have any word completion, correction or prediction tools for Marathi. Swarachakra was originally designed in 2010. It was launched on the Android platform in June 2013. Currently (September 2014), Swarachakra is available for 10 Indian languages (Hindi, Marathi, Gujarati, Telugu, Kannada, Malayalam, Odia, Punjabi, Bengali and Konkani). As of now, Marathi Swarachakra has the highest usage, and we believe that we have collected sufficient corpus for releasing in the public domain.

A small survey is conducted by us to find out the most frequent words used in the sentences so that it can help in the prediction of the next word

A. Experiment

A population of 20 people was selected and they were give a starting word which they had to use and complete the sentence.

This experiment helped in finding out the frequently used words in the sentence and to find out the most frequently used occur or a most likely to occur.

A group of 7 odd people were given मी as the starting word to which some sentences derived were मी काय करू, मी कुठे जाऊ, मी काय आणू, मी कुठे शोधू.So according to above sentences काय, कुठे were the frequently used words after मी.Similarly the next set of 7 individuals were given the word तुला as the starting word to which the sentances generated were तुला काय पाहिजे, तुला काय सांगू, तुला जमेल का The above sentences had काय as the frequently used words .Similarly last 6 individuals were given the initial as माझ्या कडे येशील का ,माझ्या कडे आहे,माझ्या साठी कर.In the above the frequently used words कडे as frequently used words

There are many words that are most frequently used but we have taken some of the minimal examples.

B. Applying N-gram algorithm

Natural Language Processing, deals with ways in which machines derives its learning from human languages. The basic input within the NLP world is something called a Corpora, which essentially is a collection of words or groups of words, within the language[6]. Even Google has its own linguistic corpora with which it achieves many of the amazing features in many of its products. Deriving learning out of the corpora is the essence of NLP. In the context which we are discussing, i.e. word prediction, its about learning from the corpora to do prediction. The way we do learning from the corpora is through the use of some simple rules in probabilities. It all starts with calculating the frequencies of words or group of words within the corpora. For finding the frequencies, what we use is something called a n-gram model, where the "n" stands for the number of words which are grouped together. The most common n-gram models are the trigram and the bigram models. For example the sentence माझ्या कडे येशील का has following biagrams माझ्या कडे,कडे येशील,येशील का. Similarly a tri gram model will split a given sentence into combinations of three word groups. These groups of trigrams or bigrams forms the basic building blocks for calculating the frequencies of word combinations. The idea behind calculation of frequencies of word groups goes like this. Suppose we want to find the frequency of bigram माझ्या कडे What we look for in this calculation is how often we find the combination of the words "माझ्या" followed by "कडे" within the whole corpora. Suppose in our corpora there were other 5 instances where the words "माझ्या" followed by the word "कडे" , then the frequency of this bigram is 5.

Once the frequencies of the words are found, the next step is to calculate the probabilities of the bigram. The probability is just the frequency divided by the total number of bigrams within the corpora.Suppose there are around 500,000 bigrams in our corpora, then the probability of our bigram "माझ्या कडे" will be 5/500,000.The probabilities so calculated comes under a subjective probability model called the Hidden Markov Model(HMM).By the term subjective probability what we mean is the probability of an event happening subject to something else happening. In our bigram model context it means,the probability of seeing the word "कडे" subject to having preceded with words "माझ्या" Extending the same concept to bigrams, it would mean probability of seeing the

third word subject to have seen the first two word. So if "हो आहे" is a bigram, then the subjective probability would be the probability of seeing the word "आहे" followed by the word "हो" .

The trigrams and bigrams along with the calculated probabilities arranged in a huge table forms the basis of the word prediction algorithm.The mechanism of prediction works like this. Suppose you were planning to type "हो तिथे आहे" and you typed the first word "हो ". The algorithm will quickly go through the n-gram table and identify those n-grams starting with word "हो" in the order of its probabilities. So if the top words in the n-gram table starting with "हो" are हो तिथे आहे ,हो मी करते,हो तीच ती in decreasing order of probabilities, the algorithm will predict the words "आहे" ," मी" and "तीच " as your three choices as soon as you type the first word "हो" .After you type "हो" you also type "तिथे" the algorithm reworks the prediction and looks at the highest probabilities of n-gram combinations preceded with words "हो" and "तिथे" . In this case the word "आहे" might be the most probable choice which is predicted. The algorithm will keep on giving prediction as you keep on typing more and more words. At every instance of your texting process the algorithm will look at the penultimate two words you have already typed to do the prediction of the running word and the process continues.

CONCLUSION

Although prediction and gesture-based input are studied extensively for English virtual keyboards, these are relatively new features in Indian language text input and very little literature exists about them. Our work highlights some of the challenges that have to be overcome to make keyboards effective.

It is observed that autocorrection feature some times turns out to be a wrong as there are widely available words which have the same start prefixes. Provision can be made that will help to overcome the problem adding the frequently used words in the corpus so that the software which is created will able to recognize and allow to predict the correct words.

To a large extent virtual touch-screen keyboards seem to have resolved the "puzzle" of text input in Indian languages. Now the battle seems to have moved on to the frontier of speed.

IV. FUTURE WORK

The idea is to create a mechanism which will help to predict the next words as well as help in autocorrecting the Marathi text input for the smartphone users.There are difficulties faced for those people who might not be aware about the correct phrase or the next word to be used in such a scenario it is very helpful if we have a corpus which might help in creating a solution which can be further enhanced for the other Indian languages.This small research is to suggest the inclusion of this features to the swarachakra which will enhance its features.

In this paper we have presented a preliminary analysis of the corpus conducted by the developers of the swarachakra,we have tried to explain what is the future enhancement which can be done in the swarachakra so that it becomes flexible to a non Marathi person as well to input the text with ease.We acknowledge that a lot remains to be done.The problems face in the inputting text has already been seen and the reson might be variation may have occurred either to save space / keystrokes, due to a change of language, or for stylistic reasons. Identifying this may require both quantitative as well as qualitative studies as well as linguistic expertise. It will also be interesting to compare differences between conversational and formal Marathi with other languages (e.g. English, Hindi etc.).

.

REFERENCES

[1] https://en.wikipedia.org/wiki/Predictive_text

[2] Baker P., B. K., Corpus linguistics and South Asian languages: Corpus creation and tool development. Literary and Linguistic Computing 19, 4 (2004), 509-524.

[3] Forum for Information Retrieval (FIRE), Information Retrieval Society of India. (12 2-4). Mumbai, Maharashtra, India. Retrieved from http://www.isical.ac.in/~fire/2011/index.html

[4] Wikimedia Downloads. (n.d.). Retrieved 07 07, from Wikimedia: http://dumps.wikimedia.org/mrwikisource/20140705/, 2014.

[5] .Joshi, A., Dalvi, G., Joshi, M., Rashinkar, P., Sarangdhar, A. Design and evaluation of Devanagari virtual keyboards for touch screen mobile phones. MobileHCI 2011, 323- 332.

[6] https://bayesianquest.wordpress.com/2015/12/06/machine-learning-in-action-word-prediction/