Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 2; January-March, 2015 pp. 177-180

http://www.krishisanskriti.org/acsit.html

Word Prediction Systems: A Survey

Riya Makkar[1], Manjinder Kaur[2] and Dharam Veer Sharma[3]

[1]Research Student, M.Tech(CSE) Department of Computer Science, Punjabi University
[2]Research Student, M.Tech(CSE) Department of Computer Science, Punjabi University, Patiala
[3]Department of Computer Science, Punjabi University, Patiala
E-mail: [1]patialariyamakkar25@gmail.com, [2]manjinder.91@gmail.com, [3]dveer72@hotmail.com

Abstract—*Information technology plays a very important role in society. Peoples with disabilities are often limited by slow text input speed despite the use of assistive devices. Word Prediction aims at easing the text entry by offering the next word or by suggesting the list of most probable words.Many word prediction methods can be found in literature,which are used to implement various word prediction systems in different languages. This heterogeneity makes it difficult for the user to understand, compare and select the most convenient prediction system. This paper gives an overview of various prediction methodsand presents a survey on word prediction systems that implements these methods for different languages. The efficiency for various predictors is examined and the results are compared.*

1. INTRODUCTION

Many peoples in the world suffer from physical, perceptive and/or cognitive disabilities and are slow typists.So, in order to help those peoples several assistance technologies have been developed such as Word Prediction.

Word prediction is the problem of guessing which words are likely tofollow a given segment of text or a sequence of words. A computer program performing word prediction is called a word predictor [1], [2].The system typically works by predicting the most probable characters or words for the current position of the sentence typed by the user. As the user uses the program over time, the words that the user uses more commonly are predicted more frequently.Then, the system updates the list accordingto the sequence of the so-far entered letters. Next, a list ofthemost common words or phrases that could come after theselected word would appear. The process continues until thetext is completed [3].

Word Prediction implies both 'Word Completion' and 'Word prediction' to increase the text Production rate.Word completion deals with suggesting the user a list of words after a letter has been typed, while Word prediction deals with suggesting the user a list of probable words after a word has been typed or selected, based on previous words rather than on the basis of the letter.

The goal of all writing assistance systems is to enhance the Keystroke Saving Rate (KSR) which is the number of keystrokes that the user saves by using the word prediction system. Higher the value of KSRbetteris the performance; as a result, reduces both time and effort for producing a text. Typing in a word prediction systems typically requires smaller setof keys than ordinary typing, which is useful for the peoples with physical impairments.

There are manyword prediction systems that were developed and are developing with different approaches for different languages. In this paper, various approaches towards word prediction will be discussed; also we will describe and evaluate the prediction systems for different languages.

2. APPROACHES TOWARDS WORD PREDICTION

There are many approaches developed towards word prediction that are used to model the natural language since the early 1980s.These approaches can be classified into three groups.

2.1. Statistical Modeling

In statistical prediction, the choice of words for placement in the prediction list is based upon the probability that they will appear in the text [2]. Therefore, it is also known as Probabilistic Modelling. Statistical word prediction is made based on the Markov assumption in which only the last n-1 words of the history affects the next word [3]. Therefore, the model is also called n-gram Markov model. The methods that are commonly used in statistical word prediction systems are:

1. Word frequencies: This method involves sorting the complete lexicon into their frequency order, and offers the few at the top of the list to user as predictions. In other words, the systems used unigram word model with a fixed lexicon [3].Statistical NLP is always based on corpora because word frequencies are

calculated from this corpus therefore for training and testing the corpus containing approximately 3 million token are used [4].

2. WordSequence frequencies: This method involves the history as a clue for appearance of the next words. If only the previous word was used to predict the next word in the current position of the sentence being typed, then it was named bigram word model or first order Markov model. If the last two words were used to predict the next word, then it was named trigram word model or second order Markov model [3].If more than two words were used as a clue to predict the next word, then it is called n-gram Markov model, where n is the number of words used in probability sequence.

It is a robust method for providing smart word suggestions, allows the user to stray away from the rules of grammar therefore serves to save time.

2.2. Knowledge-Based Modeling

The systems that used statistical modeling for prediction often predict the words that are grammatically inappropriate. And therefore, impose a heavy cognition load on the user to choose the intended word and as a result the writing rate decreases. Knowledge Based Modeling involves omitting inappropriate words from the prediction list and gives more accurate results to the user.

1. Syntactic Prediction: Considering part of speech tags, and phrase structures, syntactic prediction is to ensure that the system tries to suggest grammatically appropriate words to the user. Almost all human-discourse languages are defined and structured. If one follows the grammar rules, he can be able to predict with some degree of accuracy at least the type of words what will come next. Primarily, syntactic prediction needs grammar detailing the structure of the sentences being created in order to make choices about the types of words it can offer[2].

2. Semantic Prediction: Some of the predicted words in the prediction list could be wrong semantically even though they are syntactically right. So, suggesting the words that are syntactically and semantically correct would increase the accuracy of the predictions [3]. It can be used by assigning categories to words and finding a set of rules which constrain the possible candidates for the next word [5]. This method is not widely used in word prediction, mostly because it requires complex hand coding or may be time consuming.

2.3. Heuristic Modeling (Adaptation)

Adaptive modeling is one, which adapts the system according to the user and makes appropriate predictions.It involves altering the frequency tags attached to the words contained in the corpus as the user constructs the sentences. It also includes the recency tags. This concept means a word that has already occurred in a text will be given a higher probability of use, thus, more likely to be used in that text again. The advantage of this modeling is its ability to adapt to user's requirements. But its disadvantage is that no reference is made to the grammatical structure of the sentence when making prediction.

3. ANALYSIS OF THE PREDICTION SYSTEMS FOR DIFFERENT LANGUAGES

The word prediction system as a typing aid considers its efficiency, the main aim to reduce effort and message-elaboration time. If it is included in aids for people to reduce the effort needed to write, it is necessary to decrease the number of keystrokes for composing a message. If it is included in aids for people to reduce the number of characters produced on a time unit, number of characters incorporated into the text by means of a single prediction should be larger than the number of characters written by a single selection [6].

As an example, the results shown below of the experiments with different methods, lexicons and languages are summarized in the following tables (the last column shows the relative improvement between experiment *i* and experiment *i-1*). For each language, several experiments have been run: a basic experiment (exp. 1) with the main lexicon and the grammatical information (when available), experiment 2 adding n-grams and experiment 3 adding the personal lexicon to adapt to the new vocabulary in a given topic [7].

Table 1: Results of the English word prediction system: % of saved keystrokes

Exp.	Configuration	Result	Relative Impr.
1	Static bipos and tripos	28.2%	
2	Exp. 1 plus 2-grams to 6-grams	37.4%	32.6%
3	Exp. 2 plus personal lexicon	47.7%	27.5%

Table 2: Results of the Spanish word prediction system: % of saved keystrokes

Exp.	Configuration	Result	Relative Impr.
1	Static bipos and tripos and features management	42.7%	
2	Exp. 1 plus 2-grams to 6-grams	51.9%	21.5%
3	Exp. 2 plus personal lexicon	53.3%	2.7%

Table 3: Results of the Swedish word prediction system: % of saved keystrokes

Exp.	Configuration	Result	Relative Impr.
1	Unigrams	33.8%	
2	Exp. 1 plus 2-grams to 6-grams	42.7%	26.3%
3	Exp. 2 plus personal lexicon	47.7%	11.7%

The differences in the results of experiment 1(from 28.2% for English to 42.7% in Spanish) are due to the agreement between test and training texts and the amount of information sources available for particular language.

Experiment2 depends on the agreement between the test text and the text used to train the n-grams. It always improves the results.

Experiment 3 shows how powerful the personal lexicon can be.

3.1 Influence of Indian languages in prediction

Indian languages are morphologically rich and considered as highly inflected language than English. Indian languages uses large number of base characters and there are a lot of phonetically or a graphically similar character which needs more time to search the desired characters and occasionally leads to taping wrong characters. These problems leads to develop a user friendly word prediction system augmented with virtual keyboard in the context of Indian languages.

3.1.1. Prediction results

This section presents a number of relevant results found in the literature [6].

Firstly, the result achieved with the word prediction system,***hIndiA***(as shown in [8]) are presented. It is observed that proposed system corrects 95.63% of simulated errors. It achieves 95.01% of *Hit rate* and *Keystrokes until prediction* of 1.54 in the error-free condition whereas in the presence of error, *Hit rate* is 92.25 and *Keystrokes until prediction* is 1.66. With benchmark text *H5* (In-domain data), it achieves on average, 62.52% of *Potential keystrokesavings*, 96.5% of *Hit rate*, and *Keystrokes until prediction* of 0.83. The proposed word prediction system is simple to use for both categories of users (experienced & inexperienced).

The**Lipik**, statistical predicting softwaresupports text composition in Indian languages. It has an inbuilt virtual keyboard to enter texts. The keyboard is based on the QWERTY layout. Prediction (word-level) provides ten suggestions in the prediction window. Once the word is completed, user needs to press a space bar to populate the next possible word. The scenario necessitates one additional key press on each successful completion of a word, apart from the selection of it from the prediction window [6]. The Text Entry Rate (in WPM) achieved by Lipik in Hindi, Bengali and Telugu is 5.04, 4.40, 2.76 respectively for the inexperienced users and 5.32, 4.93, 3.23 respectively for experienced users.

Google provides a word prediction mechanism to predict search keywords (in many Indian languages including Hindi). It is augmented with virtual keyboard in a QWERTY layout and displays results in the prediction window. When a user enters a prefix of a word, it returns the top ten possible suggestions in the prediction window. It also provides multiple words to be predicted and displayed in the prediction window. When the predicted word is selected from the prediction window, it searches the Internet and returns the results [6]. The Text Entry Rate (in WPM) achieved by Google in Hindi, Bengali and Telugu is 5.32, 4.64, 3.05 respectively for the inexperienced users and 5.63, 5.38, 3.35 respectively for experienced users.

Quillpad and Google Tranliteratea multi-lingual predictive transliteration system is a generic system that can be trained to predictively transliterate between any two alphabet-based languages [9]. It offers the suggestions as you type. The suggestions are ranked in the decreasing order of their statistical importance.

Table 4: Comparison of existing word prediction system in Hindi:

Metrics (unit)	Lipik	Google	*hIndiA
Keystroke Savings (%)	32.43	16.82	43.05
Prediction utilization (%)	85.72	87.12	93.84
Hit rate (%)	85.71	28.57	92.46
Text entry rate (wpm)	7.38	4.84	12.56

The systems designed and implemented by the authors of this paper are presented and evaluated in [10]. The **Urdu** virtual keyboard augmented with word prediction was evaluated on 20 students of computer science program. The average text entry speed was 13.47 wpm based on an initial two hour training of evaluation. The maximum speed achieved was 22.5 wpm. The predicted speed of text entry using this keyboard is 36.3901 wpm. With the extended training of the user the text entry speed of Urdu virtual keyboard can be improved.

The author of this paper [11] aimed at developing a low cost Bangla virtual keyboard enabled with both word and character level prediction support named as **Sulekha**. It was evaluated on the basis of the test carried out by four members of IICP (in Kolkata) who are affected by cerebral palsy and have various degrees of motor impairment. The typing rate that is the number of characters that a user can type per minute has been recorded. The rate has been measured when the users have not taken the help of prediction and also when they have taken the same.

Antaryami a smart keyboard augmented with a word predictor is evaluated for Hindi and Bengali Language. The overall performance of the final system measured on the test set of 200 sentences is found 85.43% forHindi and 88.83% for Bengali [12]. It used the half-typed word prediction and next word prediction to measure the prediction qualities for typed-in Indian languages. And accuracies have been measured with

word positions. Here is the table 5 showing the results of both prediction types and their accuracies.

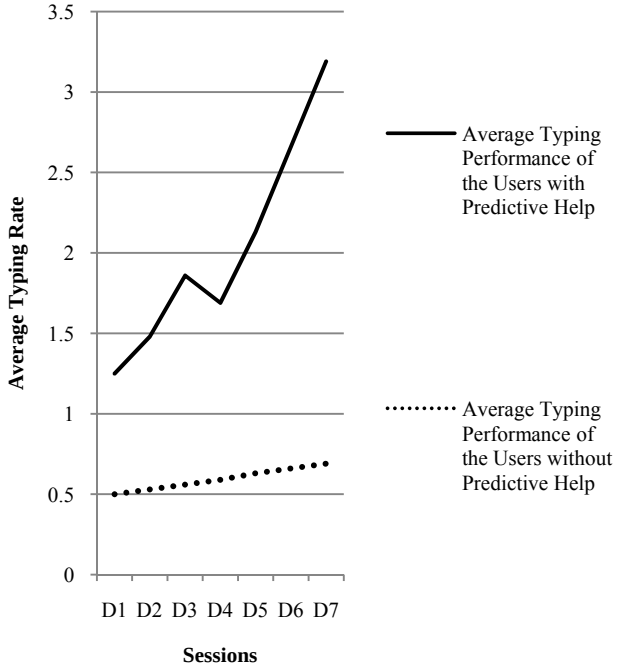

Fig. 1: Average Performance of the Users

Table 5.Word Prediction Accuracies

Prediction Types	Word Positions		
	1-3	3-6	6-9
Half-Typed word	46%	65%	68%
Next Word	52%	70%	71%
With User Modeling			
Half-Typed word	52%	75%	82%
Next Word	68%	80%	85%

4. CONCLUSION

Word prediction techniques have been frequently designed with the aim to accelerate the typing speed, increase the communication rate and to reduce the effort needed to type a text. These techniques have been initially included in aids for the people with motor disabilities, improves the quality of life, but non-disabled people can also use them while composing messages to provide more comfort and spelling assistance.

Several authors proposed various predictive modeling methods and strategies to develop a word prediction system. This paper has presented a study of these methods and analysed for various languages like English, Swedish, Spanish and Indian languages (like Hindi, Bengali, and Urdu).

The various prediction metrics and their results have been shown. Mainly the results are expressed through the Keystroke Saving in case of English, Swedish and Spanish language.

This paper also mentioned the influence of Indian languages in predicting the text, discussed the different word prediction systems and their performance measures like Keystroke Saving, Hit Rate and Text Entry Rate. The comparison of existing word prediction systems in Hindi is also discussed.

Additionally, this paper has also addressed the keyboards in different languages like Urdu, Hindi and Bengali augmented with word prediction and their overall performance is measured.

In addition to the above mentioned measures and the performance of various word prediction systems, it is better to perform usability test in order to select the most suitable system by the user.

REFERENCES

[1] J. Carlberger, “Design and Implementation of a Probabilistic Word Prediction Program,” Master’s Thesis in Computer Science, NADA, KTH, Stockholm (Sweden), 1997.

[2] E. Matthew, “Syntactic pre-processing in single-word prediction for disabled people,” Ph.D. Thesis, Univ. of Bristol, England, June 1996.

[3] M. Ghayoomi and S. Momtazi, “An Overview on the Existing Language Models for Prediction Systems as Writing Assistant Tools,” in Proc. IEEE international conference on Systems, Man and Cybernetics, San Antonio, TX, Oct 11-14 2009, pp.5083-5087.

[4] J. A. Mahar and G. Q. Memon, “ProbabilisticAnalysis of Sindhi Word Prediction using N-Grams,” Australian Journal of Basic and Applied Sciences,vol. 5, No. 5, Jan 2011, pp. 1137-1143.

[5] C. Aliprandi, N. Carmignani and P. Mancarella, “An Inflected-Sensitive Letter and Word Prediction System,” International Journal of Computing & Information Sciences, vol. 5, No. 2, August 2007, pp. 79-85.

[6] N. Garay-Vitoria, J. Abascal, “Text prediction systems: a survey,” Universal Access in the Information Society, vol. 4, Feb. 2006, pp. 188 – 203.

[7] S.E. Palazuelos-Cagigas, J.L. Martín-Sánchez, L.H. Sabatela, and J.M. Guarasa, "Design and Evaluation of a Versatile Architecture for a Multilingual Word Prediction System," inProc. ICCHP, 2006, pp.894-901.

[8] M. K. Sharma and D. Samanta, “Word prediction system for text entry in Hindi,” ACM Trans. Asian Lang. Inform. Process. vol. 13, Issue 2, Article 8, June 2014, pp. 1- 29.

[9] R. Prakash H, “Quillpad Multilingual Predictive Transliteration System,” in Proc. 24th Int. Conf on Computational Linguistics, Dec. 2012, pp. 107-114.

[10] M. Aamir Khan, M. Abid Khan and M. Naveed Ali, “Design of Urdu Virtual Keyboard,” in Proc. Conf on Language & Technology, Jan. 2009, pp. 126-130.

[11] A. Mukherjee and A. Basu, “A Virtual Predictive Keyboard as a Learning Aid for People with Neuro-Motor Disorders", in Proc. 5th IEEE International Conference on Advanced Learning Technologies, 2005, pp. 1032-1036.

[12] A. Das, “Antaryāmī: The Smart Keyboard for Indian Languages,” In the Workshop on Techniques on Basic Tool Creation and Its Applications (TBTCIA 2013), ICON,Dec. 2013.