Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018)
IEEE Xplore Part Number:CFP18AWO-ART; ISBN:978-1-5386-4765-3

Automatic Formation, Termination & Correction of Assamese word using Predictive & Syntactic NLP

Manash Pratim Bhuyan
Dept. of Information Technology
Gauhati University
Guwahati-14, India
mpratim250@gmail.com

Prof. Shikhar Kumar Sarma
Dept. of Information Technology
Gauhati University
Guwahati-14, India
sks001@gmail.com

***Abstract*—**Automatic Formation, Termination & Correction of Assamese words is a method in which a user will get the relevant suggestions for the word which he/she is intended to write. In the formation and termination method user will type a letter and the system will display the most probable words related to that letter and if the user finds required word he/she will select the word, If the required word does not appear in the suggestion list then the user has to type more letters along with the previous letter to purify the suggestion list. In the correction method, if a user type a word which is not in the corpus then the system will show one warning for that word which prevents the occurrence of the errors.

Keywords—word prediction, n-gram, probability, corpus

I. INTRODUCTION

Automatic Formation, Termination & Correction method helps the user to write in a particular natural language efficiently. This method is not only suggesting the word that the user intended to write; system is also predict the next word related to the previous word. Currently, the natural language use in this method is Assamese but the method is not language dependent from any point of view. The system can work on the corpus of any natural language where the tokens are separated by the space. This approach can help the differently abled people to write correctly and also increase the typing speed. It is a technique which can prevent the occurrence of word error. The rest of the paper is organized as follows:

Section II reflects about the related works, Section III describes the proposed model and also the procedure to calculate the probabilities of the n-gram models, Section IV analyses the results, Finally Section V concludes the the work and proposes the future work.

II. RELATED WORKS

Saharia et. al [1] have designed LuitPad Assamese writing software which can help the user to write the word by suggesting the related words from the part of a word and also suggest the letters which are mostly related to a particular letter.

Haque et al.[2] has developed a method for word prediction in Bangla language using stochastic language models and performance of the prediction system is evaluated by using unigram, bigram, trigram, backoff and deleted interpolation method.

Bickel et. al [3] derived a solution for sentence completion problem using linear interpolation of N -gram models. They derived a *k* best Viterbi decoding algorithm with a confidence-based stopping criterion which conjectures the words that most likely succeed an initial fragment.

M. Ghayoomi et. al [4] concluded that the best approach to have appropriate predictions by combinig more linguistic knowledge such as syntactic, semantic, and pragmatic in addition with the statistical knowledge at the same time to save more keystrokes. . This approach might be closer to the 100% KSS (KeyStrokes Saving).

D. C. Cavalieri [5] et. al accomplished the word prediction task by using exponential interpolation to merge POS (Part-Of Speech) based language model and a word based n-gram language model.

Spiccia [6] et. al said that word prediction generally relies on n-grams occurrence statistics, which may have huge data storage requirements and does not take into account the general meaning of the text. A method based on LSA (Latent Semantic Analysis), to resolve these issues had been proposed. An asymmetric Word-Word frequency matrix was employed to achieve higher scalability with large training datasets than the classic Word-Document approach. They also proposed a function for scoring candidate terms for the missing word in a sentence. How this function approximates the probability of occurrence of a given candidate word had been shown. Experimentally found that the proposed method outperforms non neural network language models.

C. Aliprandi et. al [7] had evaluated FastType performance enhancements for an Italian language, which is a inflected language. C. Aliprandi et. al abled to achieve word prediction Keystroke Saving up to 51% for a standard prediction list of length 10. They enriched the Language Model with morpho-syntactic information and provided the prediction method with an on-the-fly Part-of-Speech word tagger and large lexicon dictionaries.

Q. Abbas et. al [8] claimed that their model help the handicapped people to type fast just like normal human being and also strengthens the normal ones further ahead. They achieved overall 65.28% of KS is comparable or better than the state of the art resources in the domain of Urdu language and also said that one was a positive contribution in Urdu language processing. Their model has a quality of boosting with the increase in length L of the text, which is quite good in case of inflected languages like Urdu.

Habib et. al [9] focused their research on modeling, training and recall techniques for automatic sentence completion using supervised machine learning technique based on popular N-gram language model. N-gram based word

prediction works well for English, but for Bangla language, it was found more challenging to get very good, e.g. more than 90% accuracy, performance as it depends on training corpus of size more than six hundred thousand sentences. Though during the several phases of experiments, the top three models Trigram, Backoff Linear Interpolation showed almost same level of accuracy which was above 70%, however in terms of both accuracy and failure rate, the linear interpolation outperforms the other models.

Zagler et. al[11] the FASTY (predictive typing, enhancing text input user interface developing, empowering disabled people) would assist language impaired persons to write texts faster, with less physical/cognitive load and with better spelling and grammar. FASTY would be configurable for different types of disabilities, different communication settings and different European languages. It would allow easier access to PC based office systems, to modern forms of IT communication and a faster usage of text to speech synthesizers for voice communication.

Troiano et. al[12] in their paper, they showed how prediction could be used to optimize the UI layout on mobile devices, alongside the most common use in auto filling and suggesting forthcoming entries. The efficiency of the proposed system depends on the quality of predictions and fulfillment of the expectations of the user. From a panel of users survey showed that efficiency of the proposed system can be fully achieved.

III. Proposed model

A system is designed which can predict words after a letter entered by the user. The process will do corpus look up every time after each input by the user. Each time the predictor returns top five or six word related to his/her input and if the user wants to use any of the input he/she will select the predicted word or he/she can skip the prediction if the prediction is unable to meet his/her expectation and he/she type the next letter. Figure 1 shows the proposed working model.

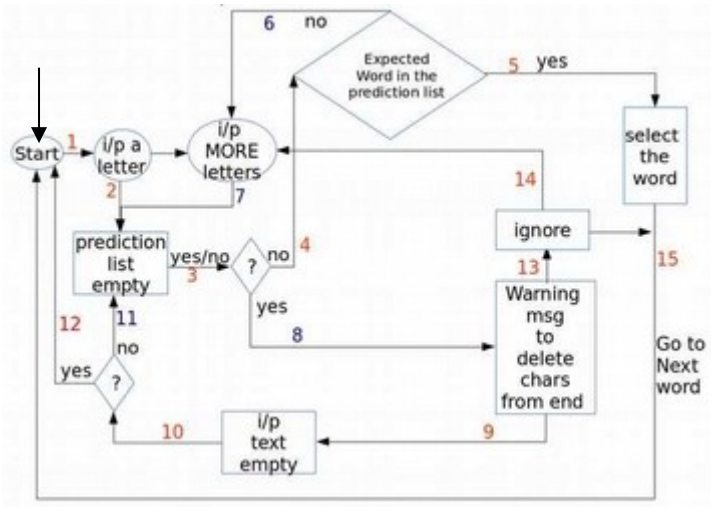

Figure 1: Word Formation, Termination and Correction model

In the Figure 1 the labeled on the edges indicates the step numbers. There are around 5 cycles in the Figure for word formation, termination and correction. If the expected word is formed in the first attempt the process will terminate and go for the next word, otherwise to bring the expected word the user has to enter more letters of that word. For error correction, if a user enters letters for which no prediction appears then warning message is displayed indicating the word to be an erroneous word, warning message request the user to delete the letters of the current word one by one until a prediction appears in the suggestion list. The warning message can be ignored, because the system will allow a user to write the name of a person, place or any other entity which is valid but not in the corpus.

A. N-gram Model:

N-gram language model is a probabilistic model where the approximate matching of next item is very high. We select N-gram based word prediction method because these are more statistical approach and accurate to predict the next word in a sentence and N-gram language model is a natural approach to the construction of sentence completion in a system.

Example: "১৩ শতিকাৰ পৰা ঊনৈশ শতিকাৰ প্ৰথমছোৱালৈকে অসমত আহোমসকলে ৰাজত্ব ___”

In the blank space the proper word is "কৰিছিল”, which we can guess by observing the previous sequence of the sentence, this is how the proposed model works.

Types of N-gram are:-

1. Unigram (1-gram) model
2. Bigram (2-gram) model
3. Trigram (3-gram) model
4. Quadrigram (4-gram) model, etc.

Sample sentence as example: "এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে |”

1. Unigram (1-gram) model:

Uni-grams of the above sentence are:

[এনে, কেচত, পুলিচে, আদালতৰ, আদেশ, বা, পৰোৱানা, নোহোৱাকৈ, আচামীক, গ্ৰেপ্তাৰ, কৰিব, পাৰে]

In unigram model each token of the sentence is a uni-gram.

There are total 12 unigrams

2. Bigram (2-gram) model:

Bi-grams of the above sentence are:

[(এনে কেচত), (কেচত পুলিচে), (পুলিচে আদালতৰ), (আদালতৰ আদেশ), (আদেশ বা),(বা পৰোৱানা), (পৰোৱানা নোহোৱাকৈ), (নোহোৱাকৈ আচামীক), (আচামীক গ্ৰেপ্তাৰ), (গ্ৰেপ্তাৰ কৰিব), (কৰিব পাৰে)]

In Bi-gram model group of two tokens of the sentence is a bi gram.

There are total 11 bigrams

3. Trigram (3-gram) model:

Tri-grams of the above sentence are:

[(এনে কেচত পুলিচে), (কেচত পুলিচে আদালতৰ), (পুলিচে আদালতৰ আদেশ), (আদালতৰ আদেশ বা), (আদেশ বা পৰোৱানা),(বা পৰোৱানা নোহোৱাকৈ), (পৰোৱানা নোহোৱাকৈ আচামীক), (নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ), (আচামীক গ্ৰেপ্তাৰ কৰিব), (গ্ৰেপ্তাৰ কৰিব পাৰে)]

In Tri-gram model group of two tokens of the sentence is a tri-gram. There are total 10 trigrams

4. Quadrigram (4-gram) model:

Quadrigrams of the above sentence are:

[(এনে কেচত পুলিচে আদালতৰ), (কেচত পুলিচে আদালতৰ আদেশ), (পুলিচে আদালতৰ আদেশ বা), (আদালতৰ আদেশ বা পৰোৱানা), (আদেশ বা পৰোৱানা নোহোৱাকৈ),(বা পৰোৱানা নোহোৱাকৈ আচামীক), (পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ), (নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব), (আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে)]

In Quadrigram model group of two tokens of the sentence is a quadri-gram. There are total 9 quadrigrams

B. Probability calculation in the various N-gram models:

Sample sentence as example: "এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে |"

Corpus size = 51, 447

Unigram (1-gram) model:

P(এনে) = 8.289E-3, P(কেচত) = 0.667E-3, P(পুলিচে) = 2.477E-3, P(আদালতৰ) =2.477E-3, P(আদেশ) = 2.191E-3 ,P(বা)= 42.59E-3, P(পৰোৱানা) = 4E-3, P(নোহোৱাকৈ) = 1.143E-3,

P(আচামীক) = 0.953E-3,P(গ্ৰেপ্তাৰ) = 6.955E-3, P(কৰিব) = 34.7E-3, P(পাৰে) = 13.24E-3

P(sentence=এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে) =

P(এনে) x P(কেচত) x P(পুলিচে) x P(আদালতৰ) x P(আদেশ) x P(বা) x P(পৰোৱানা) x P(নোহোৱাকৈ) x P(আচামীক) x P(গ্ৰেপ্তাৰ) x P(কৰিব) x P(পাৰে)

= 8.289E-3 x 0.667E-3 x 2.477E-3 x 2.477E-3 x 2.191E-3 x 42.59E-3 x 4.0E-3 x 1.143E-3 x 0.953E-3 x 0.6955E-3 x 34.7E-3 x 13.24E-3

= 4.4E-29

Bigram (2-gram) model:

P(এনে) = 8.289E-3, P(কেচত | এনে) = 0.0133, P(পুলিচে | কেচত) = 0.184, P(আদালতৰ | পুলিচে) = 0.056, P(আদেশ | আদালতৰ) = 0.073, P(বা | আদেশ) = 0.106, P(পৰোৱানা | বা) = 0.003, P(নোহোৱাকৈ | পৰোৱানা) = 0.051, P(আচামীক | নোহোৱাকৈ) = 0.108, P(গ্ৰেপ্তাৰ | আচামীক) = 0.15, P(কৰিব | গ্ৰেপ্তাৰ) = 0.067 P(পাৰে | কৰিব) = 0.016

P(Sentence = এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে)=

P(এনে) x P(কেচত | এনে) x P(পুলিচে | কেচত) x P(আদালতৰ | পুলিচে) x P(আদেশ | আদালতৰ) x P(বা | আদেশ) x P(পৰোৱানা | বা) x P(নোহোৱাকৈ | পৰোৱানা) x P(আচামীক | নোহোৱাকৈ) x P(গ্ৰেপ্তাৰ | আচামীক) x P(কৰিব | গ্ৰেপ্তাৰ) x P(পাৰে | কৰিব)

= 8.289E-3 x 0.0133 x 0.184 x 0.056 x 0.073 x 0.106 x 0.003 x 0.051 x 0.108 x 0.15 x 0.067 x 0.016

= 2.4E-15

Trigram (3-gram) model:

Sentence = এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে

P(এনে) = 8.289E-3, P(কেচত | এনে) = 0.0133,

P(পুলিচে | এনে কেচত) = 1, P(আদালতৰ | কেচত পুলিচে) = 1,

P(আদেশ | পুলিচে আদালতৰ) = 0.56,

P(বা | আদালতৰ আদেশ) = 0.61, P(পৰোৱানা | আদেশ বা) = 0.61

P(নোহোৱাকৈ | বা পৰোৱানা) = 0.393,

P(আচামীক | পৰোৱানা নোহোৱাকৈ) = 0.28,

P(গ্ৰেপ্তাৰ | নোহোৱাকৈ আচামীক) = 1

P(কৰিব | আচামীক গ্ৰেপ্তাৰ) = 1, P(পাৰে | গ্ৰেপ্তাৰ কৰিব) = 0.196

P(Sentence = এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে)=

P(এনে) x P(কেচত | এনে) x P(পুলিচে | এনে কেচত) x

P(আদালতৰ | কেচত পুলিচে) x P(আদেশ | পুলিচে আদালতৰ) x

P(বা | আদালতৰ আদেশ)x P(পৰোৱানা | আদেশ বা) x P(নোহোৱাকৈ | বা পৰোৱানা) x P(আচামীক | পৰোৱানা নোহোৱাকৈ) x P(গ্ৰেপ্তাৰ | নোহোৱাকৈ আচামীক) x P(কৰিব | আচামীক গ্ৰেপ্তাৰ) x P(পাৰে | গ্ৰেপ্তাৰ কৰিব)

= 8.289E-3 x 0.0133 x 1 x 1 x 0.56 x 0.61 x 0.61 x 0.393 x 0.28 x 1 x 1 x 0.196

= 5E-7

Quadrigram (4-gram) model:

P(এনে) = 8.289E-3, P(কেচত | এনে) = 0.0133,

P(পুলিচে | এনে কেচত) = 1, P(আদালতৰ | এনে কেচত পুলিচে) = 1,

P(আদেশ | কেচত পুলিচে আদালতৰ) = 1,

P(বা | পুলিচে আদালতৰ আদেশ) =1,

P(পৰোৱানা | আদালতৰ আদেশ বা) = 0.44,

P(নোহোৱাকৈ | আদেশ বা পৰোৱানা) = 0.44,

P(আচামীক | বা পৰোৱানা নোহোৱাকৈ) = 0.22,

P(গ্ৰেপ্তাৰ | পৰোৱানা নোহোৱাকৈ আচামীক) = 1,

P(কৰিব | নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ) = 1,

P(পাৰে | আচামীক গ্ৰেপ্তাৰ কৰিব) = 0.33

P(Sentence = এনে কেচত পুলিচে আদালতৰ আদেশ বা পৰোৱানা নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ কৰিব পাৰে)

= P(এনে) x P(কেচত | এনে) x P(পুলিচে | এনে কেচত) x

P(আদালতৰ | এনে কেচত পুলিচে) x P(আদেশ | কেচত পুলিচে আদালতৰ) x

P(বা | পুলিচে আদালতৰ আদেশ) x P(পৰোৱানা | আদালতৰ আদেশ বা) x

P(নোহোৱাকৈ | আদেশ বা পৰোৱানা) x P(আচামীক | বা পৰোৱানা নোহোৱাকৈ) x P(গ্ৰেপ্তাৰ | পৰোৱানা নোহোৱাকৈ আচামীক) x P(কৰিব | নোহোৱাকৈ আচামীক গ্ৰেপ্তাৰ) x P(পাৰে | আচামীক গ্ৰেপ্তাৰ কৰিব)

= 8.289E-3 x 0.0133 x 1 x 1 x 1 x 1 x 0.44 x 0.44 x 0.22 x 1 x 1 x 0.33

= 1.5E-6

IV. RESULT AND DISCUSSIONS

One experiment is performed to compare the various n-gram models so that in the future experiments the best n-gram model can be used to get the efficient result. For eight different sentences the probabilities are calculated in the four n-gram models which are shown in the Table 1.

Table 1: Probability of the sentences in the various n-gram models

Sentence	No. of words	Probability			
		Uni-gram	Bi-gram	Tri-gram	Quadri-gram
S1	5	2.60E-13	4.10E-08	6.50E-05	2.80E-04
S2	7	2.05E-21	4.00E-08	4.40E-04	4.40E-04
S3	12	4.40E-29	2.40E-15	5.00E-07	1.50E-06
S4	16	2.70E-48	2.90E-19	1.81E-05	1.45E-05
S5	18	1.86E-46	5.40E-26	1.80E-07	1.32E-04
S6	23	9.00E-64	3.95E-34	2.58E-08	4.75E-05
S7	28	1.76E-79	2.82E-33	2.35E-06	9.90E-05
S8	31	3.54E-94	4.00E-30	1.00E-07	9.00E-05
Average		**3.25E-14**	**1.01E-08**	**6.58E-05**	**1.38E-04**

The average sentence probabilities are shown in the Figure 2.

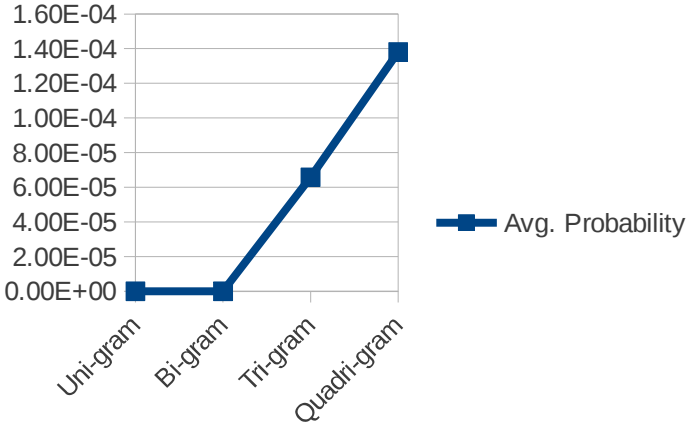

Figure 2: Avg. Sentence probability graph in various N-gram models

From the above Figure 2 it is seen that the sentence probabilities in trigram and quadrigram models are not very much deviated from each other. The testing environment for the experiment is shown in the Figure 3

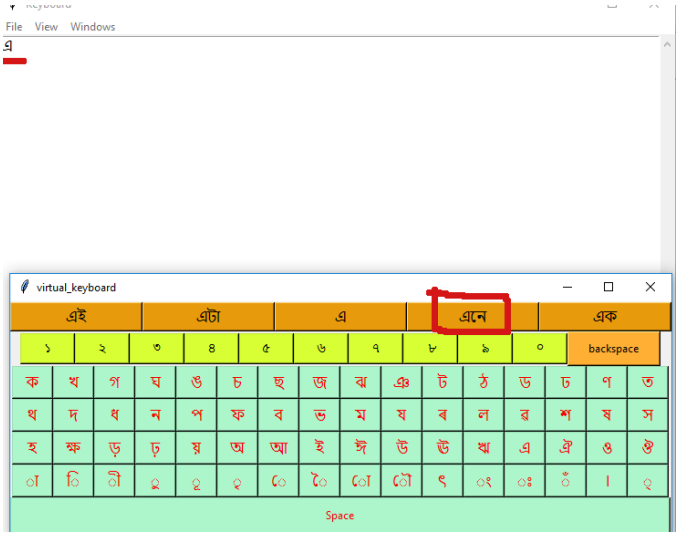

Figure 3: Snapshot of the testing GUI

V. CONCLUSION AND FUTURE WORK

Automatic word formation, termination and correction system can help people to increase their writing speed by predicting the relevant words. Among the various n-gram models after comparing them from the experimental result it is found that trigram and quadrigram models showing almost similar kind of result, but at the same time the corpus size for trigram and quadrigram has a significant difference. So, from the above experiment we can conclude that for relativly faster computer a user may go for the quadrigram model and for economical computer user may confined in trigram or bigram. In addition the proposed model also contains syntactic level prediction because the proposed model stores the fragmented sentences which are syntactically correct.

In the future linear interpolation model can incorporated to check the performance of the system. In addition, accuracy and the efficiency of the proposed model will be evaluated with a group of users and their feedback. Corpus size can be increased to get more accurate results. To enhance the n-grams, syntactic n-gram can be implemented to reduce the size of the n-gram corpus.

REFERENCES

[1] N. Saharia and K. M. Konwar, “LuitPad: a fully unicode compatible Assamese writing software,” in, Proceedings of the 2nd Workshop on Advances in Text Input Methods (WTIM 2), Mumbai, India, 2012, pp. 79-88.

[2] M. Haque, M. T. Habib, M. M. Rahman, "Automated word prediction in Bangla language using stochastic language model", *International Journal in fundation Computer Science& Technology (IJFCST)*, vol. 5, Nov. 2015.

[3] S. Bickel, P. Haider, and T. Scheffer. 2005. Predicting sentences using n-gram language models. In Proceed-ings of the conference on Human

Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 193–200, Stroudsburg, PA, USA. Association for Computational Linguistics.

[4] M. Ghayoomi and S. Momtazi, "An overview on the existing language models for prediction systems as writing assistant tools," in Systems,

Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, San Antonio, Texas, 11-14 October 2009, pp. 5083–5087, iSSN: 1062-922X.

[5] D. C. Cavalieri, S. E. Palazuelos-Cagigas, T. F. Bastos-Filho, and M. Sarcinelli-Filho, "Combination of language models for word prediction: an exponential approach," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, pp. 1481–1494, 2016.

[6] C. Spiccia, A. Augello, G. Pilato and G. Vassallo, "A word prediction methodology for automatic sentence completion," *2015 IEEE International Conference on Semantic Computing (ICSC)*, Anaheim, CA, USA, 2015, pp. 240-243. doi:10.1109/ICOSC.2015.7050813

[7] C. Aliprandi, N. Carmignani, N. Deha, P. Mancarella, and M. Rubino, "Advances in nlp applied to word prediction," 2008.

[8] Q. Abbas, (2014), "A Stochastic Prediction Interface for Urdu", Intelligent Systems and Applications, Vol.7, No.1, pp 94-100 .

[9] Habib, Md & AL-Mamun, Abdullah & Rahman, Md & Md Tanvir Siddiquee, Shah & Ahmed, Farruk. (2018). An Exploratory Approach to Find a Novel Metric Based Optimum Language Model for Automatic Bangla Word Prediction. International Journal of Intelligent Systems and Applications. 2. 47-54. 10.5815/ijisa.2018.02.05.

[10] Al-Mubaid, H.: A Learning - Classification based Appro Word Prediction. The International Arab Journal of Information Technology 4(3), 264–271 (2007)

[11] W.L. Zagler, C. Beck, G. Seisenbacher: "FASTY - Faster and easier text generation for disabled people"; Presentation: AAATE '03, Dublin; 08-30-2003 - 09-03-2003; in: "Assistive Technology - Shaping the Future"; G. Craddock, L. McCo rmack, R. Reilly, H. Knops (ed.); IOS Press, Volume 11 (2003), 1 58603 373 5; 964 - 968.

[12] *Troiano, L., Birtolo, C., & Armenise, R. (2017). Modeling and predicting the user next input by Bayesian reasoning. Soft Comput., 21, 1583-1600.*