March- April 2020
ISSN: 0193-4120 Page No. 1169 - 1173

Next Word Prediction using N-gram model and Smoothing

Alok Pandey
CDAC ATC Netcom,Jaipur

Tanu Jain
Department of Computer Science, Banasthali Vidyapith
tanujain1233@gmail.com

Drashika Khanna
Department of Computer Science, Banasthali Vidyapith

Shreeya Johari
Department of Computer Science, Banasthali Vidyapith

Upasana Khare
Department of Computer Science, Banasthali Vidyapith

Article Info
Volume 83
Page Number: 1169 - 1173
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 14 March 2020*

Abstract
Word prediction is an intelligent word processing feature that can help students and other persons simply by reducing the number of keystrokes required for typing text. The Natural Language Processing (NLP) based n-gram model is a statistical language model for fast messaging. This model is used to assign probabilities to sequences and sentences. This methodology is sort of tough in alternative languages with the exception of English. The model goes through the corpus of English words and predicts the next word using n-gram. This includes hunting numerous reviews and analysis papers concerning completely different languages, corpus and prediction techniques, and the way these prediction techniques will facilitate us in the next word prediction application. It is becoming more and more important because of the key role that it plays in various domains. This is often a difficult task since the word presents a high variance.

Keywords*: Corpus, N-Gram, Natural Language Toolkit, Perplexity, Smoothing*

I. INTRODUCTION

Mobile devices became indispensable everyday companions at home and work, to socialize, play, and do business. However, having a straightforward keyboard on bit screen devices, particularly, may be cumbersome. Automatic text prediction aims to unravel this by using previously entered text to predict the future words. Text prediction is a region of study within Natural Language Processing aims to guess the next most likely words found within the previous context [1]. The most widely understood use case for this is the most likely upcoming text features found in smartphones that permit the users to pick out future presumably word rather than writing it themselves. Text prediction algorithms which are widely used for automatic sentence

completion are immensely mentioned in the literature.

Corpus plays a key role within the field of NLP. NLTK (Natural Language Toolkit) provides various sets of corpora. There are tasks such as spelling error detection, word prediction for which the location of the punctuation is important. Our model counts punctuation as words. Moreover, this analysis was performed on all ascertained single words, 2-word combinations (bigrams) and 3-word combinations (trigrams). The model provides the power to auto-complete words and suggests predictions for future words [5]. This makes writing quicker, additional intelligent and reduces effort. One feature of this model is the prediction is this fast that future word is searched as shortly as you write one word. No "predict" button is needed to create the formula operate. The main objective of the paper includes:

We aim to utilize natural language processing algorithms to predict the next word. The goal is to use computational efficient algorithms with high accuracy. The user desires to write the text while not having much information about vocabulary and conjointly wish to write in a very skilled manner too with a much better Interface.

The paper is organized into four sections where first section is introduction and next subsequent sections are related work, methodology, results, conclusion and future work respectively.

II. RELATED WORK

Among the most influential work in The Next Word Prediction can be attributed to S. Rajaraajeswari [6]. They have found that the rising Bag of words (BOW) model looks to be a far better fit predicting one thing by process text content. In this approach the model trained by the corpus and do the tokenization of the words and after this task it creates the vocabulary which also represent the repetition of the words or word frequency. The complete model is work as the word classifier. These words selected from the corpus act as a training data for the model.

Some of the most useful work was done by Shashi Pal Singh [7] they have also worked on NLP and do the next word prediction along with completing the word. In this work if we are writing some words in our text it model will auto complete the whole word according to its learning by corpus at the same time it will also predict the next word. The N-gram technique is adopted by the author to complete this task but the accuracy of predicting the word is only 37-40%. Reference [5] worked on the NLP and predict the text. The method used by the author is N-Gram model with probabilistic approach. Suitable Predictions for new words are generated with respect to the previously written text. It assigns a probability to those words which have the chances of coming next and select the next word with the highest probability value. Reference [3] uses two models n-gram model and the Hidden Markov Model. These two models are combined along with efficiency to establish the sequence of blocks in the news story.

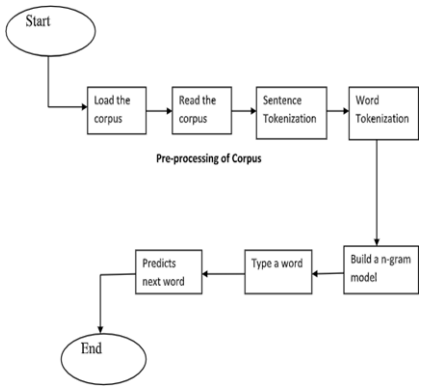

Fig.1: Methodology

III. METHODOLOGY USED

3.1 Corpus Description

The study within the field of word prediction needs a rich set of the corpus to test the algorithm for the prediction of the consequent

word. Words have completely different meanings in different contexts for that we need a large set of corpus to predict the next word with respect to the previously written text.

As with many of the statistical models in our field, the probabilities of an n-gram model come from training set the corpus it is trained on. We can measure the standard of an n-gram model by its performance on some unseen data known as the test set or test corpus. For our purpose, we have used two different corpora for training and testing.

The corpus beyond good and evil was used for predicting the next word. It is a training corpus. The testing corpus is Alice in Wonderland. It helped us in finding the accuracy and perplexity of the model.

3.2 Natural Language Processing

NLP is a subset of Artificial Intelligence in the field of Computer Science, concerning about the interactions between the computer (machine) and human. An analogy is that humans interact, perceive one another views, and respond with the acceptable answer. In NLP, this interaction, understanding, the response is formed by a personal computer rather than a person. The ultimate goal lies in enabling a computer to speak with a human and perceive human language a bit like humans do [2]. It has recently gained a lot of attention for representing and analyzing human language computationally. It also has wide applications in varied fields like machine translation, email spam detection, data extraction, summarization, etc.

3.3 Natural Language Tool Kit (NLTK)

This toolkit is one in all the foremost powerful NLP libraries that contains packages to form machines perceive human language and reply to it with an appropriate response. NLTK plays a major role in our model next word predictor [7]. With the help of this toolkit, we will use inbuilt functions rather than writing the whole code for that particular function. It also reduces the overhead of a programmer and helps in getting better results.

3.4 N-gram Language Model

This model is a type of language model which is used for predicting upcoming words in form of an *(n-1)* order Hidden Markov model. N-gram is a series of n words. Language model is a type of model which is based on calculating the probability based on the count of the sequence of words [3]. They are currently widely employed in probability, statistical NLP and data compression. The benefits of this model with larger value of n are simplicity and scalability. The next words W_n are predicted with a given context$(W_1, W_2, W_3 \ldots\ldots W_{n-1})$ calculates the probability function P which is derived through the Bayes theorem. The n-gram model is expressed by $P(W_i / W_1, W_2 \ldots\ldots W_{i-1})$. It is known as unigram model when the value of n=1. Similarly for bigram (n=2) and trigram (n=3). We will use the model in three forms i.e. Unigram, Bigram, Trigram which will predict the next words with better understanding.

3.5 Perplexity

The perplexity (PP) of a language model on a check set is that the inverse probability of the test check, normalized by the amount of words. It is an activity of however well a likelihood model predicts test information. In the context of the communication method, perplexity is a method to judge language models.

For a check set W = $(W_1, W_2, W_3 \ldots\ldots W_n)$, whenever N is the total number of words within the corpus:

$$PP(W) = P\,(W_1, W_2, W_3 \ldots\ldots W_N)^{\frac{-1}{N}}$$

$$= \sqrt[N]{\frac{1}{P\,(W_1, W_2, W_3 \ldots\ldots W_N)}} \qquad [1]$$

March- April 2020
ISSN: 0193-4120 Page No. 1169 - 1173

We can use the chain rule to expand the likelihood of W.

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(W_i / W_1, W_2, \ldots W_{i-1})}} \quad [2]$$

The higher the possibility of the word sequence, the lower the perplexity. Thus, maximizing the check set probability is comparable to minimizing perplexity.

3.6 Smoothing

Smoothing is a process in which we will add some content of the probability for those words who are not appear in the training data set, we will have to shave off a bit of probability mass from some more frequent events and give it to the events we have never seen. This modification is called smoothing. It is a strategy used to account for data scarcity where data is present in a very poor form. There are various types of smoothing: - add-1 (Laplace Smoothing), add-k smoothing, etc. We have use Laplace smoothing in our model. It is a technique used to smooth categorical data. In this phenomenon, we add one to all the bigram counts, before we normalize them into probabilities. All the counts that used to be zero will now have a count of 1; the counts of 1 will be 2, and so on.

IV. RESULTS

Text prediction is a vicinity of study inside the linguistic communication process that aims to supply systems that guesses the next most likely word given the words found within the previous context.

The word prediction model is also useful in constructing relationship between the words. This is the reason why an N-gram model is an important tool in predictive systems. The next important task for the model is to calculate the performance of the model. Here the performance measure is perplexity of the model. Perplexity is a method by which we can measure the performance of a model by implicit testing approach on the test data set. The test data is isolated from the training data. Our aim is to decrease processing time and the loading cost of the data.

There is one more technique which improves the performance of the system is back off in which if a sequence is missing or not found in a higher order N-gram model, then we will eliminate the first text and reduce the order of the model i.e. we switch on the next lower order sequence and perform the task. For the trigram model, it is necessary that the user types at least 2 words. But if those 2 words are not found in the trigram model then we back off to the bigram model. Along with predicting the next word, the model also calculates Perplexity which comes out to be 8.45.

V. CONCLUSION

Programmers and practitioners facing the choice of selecting the appropriate training corpora for statistical language models. This paper shows that how does the N-Gram model works in NLP task. By the above results we can see that as we increases the order of the model it will predicts the next word much better. The next important thing with the N-gram model is that the corpus size. As large as the corpus will predict the better results. It has been observed that a large amount of dissimilar data is more useful for language model training than a little amount of similar data set.

VI. FUTURE WORK

The ever-growing field of social media and instant electronic communication has created the requirement to style a system that could support quick, comfortable and smooth typing. There is most likely far more to be done to boost the model, at the level of architecture, computational efficiency, and taking advantage of previous information. An important priority of future research should be to improve speed-up techniques as well as ways to increase capacity without increasing training time too much.

REFERENCES

[1] Carmelo Spiccia, Agnese Augello, Giovanni Pilato and Giorgio Vassallo. (2015). A word prediction methodology for automatic sentence completion, Proceedings of the 2015 IEEE 9th International Conference (978-1-4799-7935).

[2] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin,(2006) A Neural Probabilistic Language Model, in Innovations in Machine Learning, Springer Berlin Heidelberg, 2006. pp. 137-186.

[3] Deepa Nagalavi and M. Hanumanthappa, (2014) N-gram Word Prediction Language Models to Identify the Sequence of Article Blocks in English E-Newspapers (978-1-5090-1022).

[4] Steffen Bickel, et al. (2005), Predicting Sentences using N-Gram Language Models", Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 193-200, Vancouver, October 2005.

[5] Jaysidh Dumbali, Nagaraja Rao A.(2019), Real Time Word Prediction Using N-Grams Model, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-5 March, 2019.

[6] Deepu S, Pethuru Raj and S.Rajaraajeswari (2016). A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction, International Journal of Advanced Networking & Applications (IJANA) ISSN: 0975-0282 in 2016.

[7] Shashi Pal Singh, Ajai Kumar, Daya Chand Mandad, Yasha Jadwani (2016) Word and Phrase Prediction Tool for English and Hindi language International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016.