International Core Journal of Engineering Volume 7 Issue 6, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0054

Application of N-grams in Language Model, Genomes and COVID-19 Virus

Zhengrong Tang*, Duotian Qin

College of Arts and Science, Stony Brook University, Stony Brook, New York, 11790, USA.

*Corresponding author. Email: tangzhengrong33@126.com

Abstract

In information theory, n-gram was defined as any n-long sub-sequences of consecutive tokens in a sequence. Since late 1940s, it has developed and applied in multiple fields of technology. This paper introduced the main three applications of n-grams in the prediction of English language model, genomes, and COVID-19 virus. In predicting English, entropy of n-gram was used to calculate the uncertainty on average of the next English letter when the previous N-1 English letters were known. It was also used to categorize and characterize genomes, etc. Under serious COVID-19 condition, n-gram precisely identified the origins of COVID-19 from different locations as well as presented the psychological effect of COVID-19 virus through social media. N-gram is a promising method for future use in multiple areas.

Keywords

n-grams; Entropy; Information Theory; Genome; COVID-19.

1. Introduction

N-gram refers to any n-long sub-sequences of consecutive tokens in a sequence [1]. As one of the overwhelming statistical language model algorithms, n-gram effectively counts the appearance frequency of all grams. The idea of informational entropy was firstly developed by Claude E. Shannon in late 1940s [2], which was considered as the most influential measurement in the information theory. The information theory has been practiced in many fields of technology, such as statistical language model, cryptography, neuroscience and etc. Based on Shannon's work, the communication system and networks developed effectively. Also, more and more algorithms were enlightened by Shannon's discovery [2].

Previously, N-gram and the entropy were applied in English language model, which were able to measure the predictability of English: how well could the next English alphabet in a text be predicted when the former N alphabets were known [3]. N-gram was also applied in genomes. For example, it was used in classification and clustering [1] as well as pattern mining in the whole genome sequences [4]. It was useful for predicting the genome promotors [5] and estimating the eukaryote proteomes [6]. Recently, COVID-19 pandemic event negatively impacts people and businesses globally. Scientists used N-gram to identify the origins of the genomes of COVID-19 virus from different geological places around the world [7]. What is more, N-gram was also applied in a common social media, Twitter, during COVID-19 pandemic event to reflect how people around the world follow the relevant event through news and stories.

This current paper is separated into 3 chapters, and it aims at introducing the application of N-grams in the language model, genome, COVID-19 virus and the attention in the social media around this topic.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0054

2. Body

2.1 English Language Model

In 1951, C.E. Shannon firstly defined two terms, entropy (H) and redundancy in the information theory. Entropy referred to how much information an English letter can provide on average. Redundancy was defined as the amount of constraint an English letter because of its statistical structure [3]. He showed the calculation of entropy from the statistics of English: by continually considering more and more of the statistics of English, entropy H approached to limit [3]. The equation below calculates the N-gram entropy (F_N), which measured the uncertainty on average of the next English letter when the previous N-1 English letters were known. Note that *p* in this formula is the probability [3].

$$F_N = -\sum_{i,j} p(b_i, j) log_2 p_{b_i}(j)$$

According to the above formula, as p_n=0 when n approached infinite and word rank 8727 was the critical n in English, the N-gram entropy for English was calculated as below [3].

$$F_N = -\sum_{1}^{8727} p_n log_2 p_n = 11.82\ bits\ per\ word$$

However, this number remained contentious. According to Grignetti, he recalculated the N-gram entropy for English as about 9.8 bits per word. By applying Zipf's formula which only approximately estimated the word frequency, he used a word frequency table and produced a more accurate result [8].

Shannon performed experiments to demonstrate how English could be predicted. For the 27 alphabets (26 alphabets and the space), 69% of the alphabets were guessed correctly [3]. And the errors occurred mostly at the beginning of syllables or words because of more possibilities of different letters to be guessed. Reduced text length also showed the similar results [3]. To figure out how English can be predicted depending on the previous known N letters, Shannon required the subjects to guess the text (each with 15 letters), which was randomly chosen from a book, letter by letter. He discovered that the most possible alphabet was the space when there were no known letters and the possibility was 0.182), and the probability of next guess was 0.107 when the previous guess was wrong [3]. It was important to notice that the prediction from subjects improved: The more N letters previously known, the better the prediction was. Although the reverse experiment, which required subjects to guess the alphabet before the known letter, was more difficult than the forward experiment, the results were just slightly worse [3].

For English language, Shannon discovered that total entropy was calculated as below [3].

$$\sum_{i=1}^{27} i(q_i^N - q_{i+1}^N) log_i$$

He also showed that there was a lower bound and an upper bound for F_N. They were both monotonic decreasing functions of N [3]. For the upper bound, q_i^{N+1} always majorized q_i^N and the entropy would be increased by any equalizing flow in sets of probabilities. For the lower bound, the quantity $\sum_i i(q_i - q_{i+1}) log_i$ was increased by any equalizing flow among q_i [3]. Note that the lower bound was only for the ideal predictor.

Based on Shannon's experiment on analyzing the predictability of sequence of letters in English, Bickel et al. developed sentence prediction by N-gram models [9]: they used an instance-based method as a baseline, and estimated the predictability of emails, reports, and food recipes [9]. They found out that the entropy for English Enron emails was 7.17, which was much higher than that for weather reports and food recipes. With the N-gram model, prediction length for the service center had the highest precision. In other collections (weather reports, food recipes, and enron emails), they

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0054

had much longer predictions but with a much lower precision [9]. They stated that the completion method based on N-gram predicted better on the recalling profile than index-based retrieval of sentences. Entropy was significant in determining the sentence completion and prediction [9].

2.2 Genomes

From July 2012 to March 2017, the genome databases has grown very rapidly with about 4 doubles [10]. To efficiently handle and process a large amount of genome information, scientists developed a lot of bioinformatical methods to categorize and characterize different organisms, such as neighbor joining [11] and maximum parsimony [12] to draw phylogenetic trees. However, these methods were not able to align highly plastic genomes to each other [1]. Thus, Tomović et al. developed a new N-gram-based classification and clustering technique to align genomes [1].

In their discovery, they introduced 19 new dissimilarity measures, which referred to the functions on two sets of sequences, showing the dissimilarity between these two sets [1]. Respectively, they isolated a random genome sequence, selected random 2/3 genome sequences, and extracted all genome sequences from HIV-1 and computed the values using each dissimilarity function for different N-gram length (n from 1 to 10) [1]. To measure whether their functions were good candidates, they measured the average classification accuracy. They found out that 9 out of the total 19 functions performed excellently, and groups of n larger than 5 gave the best success rate at about 99.6% [1]. But it was noticeable that smaller N-grams could also produce information that could not be reached by the larger N-grams. The other 10 functions performed much worse with the lowest success rate at about 1.7% [1]. With positive results in the classification of genome sequences, they successfully developed a clustering method based on pure statistical N-gram to hierarchically cluster the genome sequences: the genome tree was empty at start and then it was built in an increasing manner. While M represented the maximum value for the dissimilarity function of genome, at any threshold value V, all genomes that had one predecessor with M<V were categorized into the same group. The threshold value V should depend on the order of processed isolates [1]. By comparing to the previous methods PHYLIP [13] and CVTree software package [14] to obtain the clustering results of Ebola, Acelomata, Mitochondrion and Streptomyces, their N-gram based methods generated much better results.

Except for categorizing and characterizing genomes, N-gram was also applied in estimating the proteomes: King et al. developed a new N-gram-based Bayesian method, ngLOC, to predict subcellular proteomes of eukaryotes, specifically the protein sequence localization among ten different organelles [6]. Comparing to the two previous methods PSORT [15] and pTARGET [16] which were commonly used among scientists, ngLOC was able to predict more subcellular locations as well as to analyze offline. The accuracy was also higher (89%) for ngLOC than that for PSORT (72%) and pTARGET (83%) [6]. By extending ngLOC method to classify proteins from a single species, King et al. developed ngLOC-X to predict 10 different subcellular proteomes for 8 eukaryotic organisms. For fruitflies, ngLOC estimated about 28.1% of the data while ngLOC-X estimated about 38.7% of the data. The confidence score (CS) above 70 was 99.0% accurate for ngLOC while it was 99.2% accurate for ngLOC-X. The new developed N-gram-based Bayesian method ngLOC can play an important role in estimating the subcellular localization of a protein in the future.

In human genome, N-gram could help to perform the pattern mining. There are nearly three billion nucleotides with 2% of the coding regions [17]. The patterns in the human genome include satellite DNAs which are long repeating sequences [18], G-quartets which are sequences with high proportion of guanine base [19], Alu repeats which contain more complex patterns including an expressive sequence, a high cytosine and guanine base content and ending with a poly-A tail [20], etc. Identifying the patterns in a genome is time-consuming without a computational method. Ganapathiraju et al. developed Augmented Biological Language Modeling Toolkit (BLMT version 3.0) based on N-gram [4]. With this method, it was able to search for N-grams, large repeats and regular expressions in a large genome with much shorter time and a larger scale. In the future, this resource can be applied in

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0054

Next Gen Sequence and Personal Genomes Project for both pattern mining in an individual's genome or comparing different genomes [12].

2.3 COVID-19 Pandemic Event

The outbreak of coronavirus disease 2019 (COVID-19) has negatively influenced the whole world in both economics and public health. The virus results in pneumonia, lymphopenia, exhausted lymphocytes and a cytokine storm [21]. To discover more its information, Boujnouni et al. used computational techniques of machine learning and N-gram to precisely identify the origins of COVID-19 from different locations in the world [7]. In their research, 5 techniques of machine learning were used, including Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine. They also used N-gram to count the occurrences of a specific nucleotides (e.g. TGA) in a nuclear sequence (e.g. TGATGACTGATACA), which was able to create feature vectors from various genomic sequences [7]. They found out that the most possible inner-host for COVID-19 was pangolins, and the genomes of COVID-19 virus could be categorized as Alphacoronavirus 1 [7].

N-gram was not only used in finding more information, such as the inner-host, for COVID-19, but also was applied in the social media to discover how public react to the pandemic event. Both research groups, Saha et al. and Sethi et al. used Twitter as their targeted social media since it faces to most public and is designed for micro-blogging and self-expressing [22,23].

In Saha et al.'s research, they mainly focused on the psychosocial effects of COVID-19 pandemic by investigating how the public change their languages to describe their mental health and expressions through Twitter. They applied Sparse Additive Generative Model (SAGE) to distinguish N-gram between the control and experimental groups [22]. While comparing the data from March 24, 2020 to May 24, 2020 and March 24, 2019 to May 24, 2019, they found out that the mental health symptomatic expressions was increased by about 14%, and the support expressions were increased by about 5%, which both were directly related to COVID-19 pandemic [22]. Although the percentage of increase for the mental health symptomatic expressions support expression decrease along the time, which indicated that people were adapting to the condition, the study showed that public expressed their mental health concerns during the COVID-19 pandemic, and policymakers should take actions to reduce the psychosocial effects [22].

Sethi et al. aimed at discovering the sentiments from people under COVID-19 situation through Twitter [23]. They developed the model by using bi-class and multi-class, which were assigned by N-gram with cross-dataset evaluation for the machine learning techniques [23]. They got clean Tweets by choosing Tweets that involving hashtags of #COVID-19 and #coronavirus and removing the symbols #, @, case conversations and stop-words, tokenization and finally stemming [23]. The results showed that they got better results in the classifiers that were applied in the unigram and bigram feature set in the binary-class setting and in the classifiers that were applied only in the unigram feature set in the multi-class setting, which indicated that the prediction of emotions from the public through Tweets was overall accurate, at about 91%-93% [23-26].

3. Conclusion

In this paper, N-gram, a language model, was introduced, and it has been applied in multiple conditions and areas. It could be used to measure the predictability of English, classification and clustering as well as estimate the proteomes, and perform pattern mining for the genomes, and discover detailed information and psychosocial effects of COVID-19 and public's reactions to the pandemic event through social medias. A lot of alternative N-gram-based techniques, such as machine learning, can be explored. By the help of N-gram, scientists are able to predict web requests, categorize texts, and detect malicious code, etc. New N-gram-based developments are ongoing works for new functions in multiple areas in the future.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0054

References

[1] Tomović, A., P. Janičić, and V. Kešelj, n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. Computer Methods and Programs in Biomedicine, 2006. 81(2): p. 137-153.

[2] Shannon, C.E., A Mathematical Theory of Communication. Bell System Technical Journal, 1948. 27(4): p. 623-656.

[3] Shannon, C.E., Prediction and Entropy of Printed English. Bell System Technical Journal, 1951. 30(1): p. 50-64.

[4] Ganapathiraju, M.K., et al., Suite of Tools for Statistical N-Gram Language Modeling for Pattern Mining in Whole Genome Sequences. Journal of Bioinformatics and Computational Biology, 2012. 10(06).

[5] Rani, T.S. and R.S. Bapi, Analysis of n-Gram based Promoter Recognition Methods and Application to Whole Genome Promoter Prediction. In Silico Biology, 2009. 9(1,2): p. S1-S16.

[6] King, B.R. and C. Guda, ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. Genome Biology, 2007. 8(5).

[7] Boujnouni, M.E., A study and identification of COVID-19 viruses using N-grams with Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine. 2020.

[8] Grignetti, M.C., A note on the entropy of words in printed English. Information and Control, 1964. 7(3): p. 304-306.

[9] Bickel, S., P. Haider, and T. Scheffer, Predicting sentences using N-gram language models, in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05. 2005. p. 193-200.

[10] Langmead, B. and A. Nellore, Cloud computing for genomic data analysis and collaboration. Nature Reviews Genetics, 2018. 19(4): p. 208-219.

[11] Saitou, N.N.M., The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 1987. 4(4): p. 20.

[12] Camin, J.H. and R.R. Sokal, A Method for Deducing Branching Sequences in Phylogeny. Evolution, 1965. 19(3).

[13] Retief, J.D., Phylogenetic Analysis Using PHYLIP, in Bioinformatics Methods and Protocols. 1999. p. 243-258.

[14] Qi, J., H. Luo, and B. Hao, CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Research, 2004. 32(Web Server): p. W45-W47.

[15] Nakai, K. and P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends in Biochemical Sciences, 1999. 24(1): p. 34-35.

[16] Guda, C. and S. Subramaniam, TARGET: a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics, 2005. 21(21): p. 3963-3969.

[17] Bernardi, G., The Human Genome: Organization and Evolutionary History. Annual Review of Genetics, 1995. 29(1): p. 445-476.

[18] Ugarković, Đ. and M. Plohl, Variation in satellite DNA profiles—causes and effects. The EMBO Journal, 2002. 21(22): p. 5955-5959.

[19] Williamson, J.R., G-Quartet Structures in Telomeric DNA. Annual Review of Biophysics and Biomolecular Structure, 1994. 23(1): p. 703-730.

[20] Deininger, P.L. and M.A. Batzer, Alu Repeats and Human Disease. Molecular Genetics and Metabolism, 1999. 67(3): p. 183-193.

[21] Cao, X., COVID-19: immunopathology and its implications for therapy. Nature Reviews Immunology, 2020. 20(5): p. 269-270.

[22] Saha, K., et al., Psychosocial Effects of the COVID-19 Pandemic: Large-scale Quasi-Experimental Study on Social Media. Journal of Medical Internet Research, 2020. 22(11).

[23] Sethi, M., et al., Sentiment Identification in COVID-19 Specific Tweets, in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). 2020. p. 509-516.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0054

[24] Zhong, S., et al., WhatNext: a prediction system for Web requests using n-gram sequence models, in Proceedings of the First International Conference on Web Information Systems Engineering. 2000. p. 214-221.

[25] Hornik, K., et al., The textcat Package forn-Gram Based Text Categorization in R. Journal of Statistical Software, 2013. 52(6).

[26] Abou-Assaleh, T., et al., N-gram-based detection of new malicious code, in Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC 2004. 2004. p. 41-42 vol.2.