A Mathematical Model of Vocabulary Growth
Draft. Do not cite or distribute!

Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Comupter Science
University of Massachusetts, Amherst, MA 01003
lavrenko@cs.umass.edu

Abstract

We present a statistical model of vocabulary growth for applications involving large volumes of text. Vocabulary growth is modeled as repeated sampling of words from some underlying distribution. We derive general expressions for the expected number of unique words and the confidence interval around the expected value. We suggest a parametric form of word probabilities that leads to a closed form estimate of the expected vocabulary size. The proposed parametric distribution also fits empirical word frequencies better than the popular formula of Zipf. The main result is that under reasonable assumptions the vocabulary growth follows Gauss' hypergeometric function.

1 Introduction

In many text-related applications it is important to estimate how many distinct words will be observed if we are dealing with a text of size n. In Information Retrieval, it could be useful to guess how many inverted lists would be required to index a 10Gb collection. In speech recognition, vocabulary size will provide a bound on the out-of-vocabulary (OOV) rate for an hour of speech. In distributed retrieval, it is important to know how many times we should sample an unknown database to accurately estimate its language model [3]. An estimate of vocabulary growth could also be very useful in determining how much to smooth a given language model [5].

It is a common observation that vocabulary size grows sub-linearly with a total size of the dataset. Asymptotically, the number of unique words U_n in the vocabulary is often proportional to the square root of the total number of words n. Heaps [2] suggests a general form: $U_n \approx kn^b$, where k and b are parameters. Regrettably, there is no theoretical justification for this dependency. In this paper we present a statistical model that leads to a different formula for vocabulary size, based on the Gauss hypergeometric series.

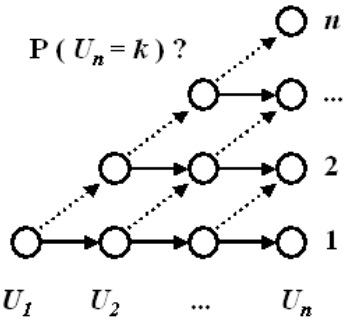

Figure 1: Vocabulary growth as a branching process: after observing $n-1$ words (U_{n-1} unique), if nth word is new we go up and have $U_n = 1 + U_{n-1}$, otherwise we stay flat.

The remainder of this paper is organized as follows. Section 2 provides detailed derivation of the dependency. Sections 2.1 and 2.2 contain very general results which are not limited to vocabulary growth. Our model for individual word probabilities in presented in Section 2.3. Section 3 presents empirical performance of our model on TREC volumes 1 and 2.

2 Theoretical Framework

2.1 Non-stationary branching process

We choose to model vocabulary growth as a sequential process. Figure 1 graphically shows the model. We consider words W_i to be arriving one-by-one in sequence from some unknown source. By U_n we denote the number of unique (distinct) words we have seen after

n words have arrived. Suppose after $n-1$ words we observed U_{n-1} unique words. When the nth word (W_n) arrives, there are two possibilities. If we have seen W_n before, our vocabulary does not grow and $U_n = U_{n-1}$ (represented by a flat transition ($\rightarrow$) in Figure 1). Otherwise, W_n is a new word, in which case our vocabulary grows by one and $U_n = 1 + U_{n-1}$ (upward transition ($\nearrow$) in the diagram).

This formalism is known as a branching process or a random walk in statistics. The process is completely determined by the probabilities of going up $P_n(\nearrow)$ and staying flat $P_n(\rightarrow)$ at every step n. If the probabilities are stationary, i.e. if $P_n(\nearrow)$ is the same for every node in the graph, the whole process can be trivially described by a binomial distribution. However, if $P_n(\nearrow)$ varies with n, it becomes exceedingly hard to estimate $P(U_n = u)$ for arbitrary n and u.

Fortunately, it is possible to get simple estimates for the expected value and the variance without computing $P(U_n = u)$. The following equations are true for any random walk where $P_n(\nearrow)$ depends on n but not on U_{n-1}. Proofs are omitted for brevity and will be included in the full paper.

$$E[U_n] = \sum_{i=1}^{n} P_i(\nearrow) \tag{1}$$

$$V[U_n] = \sum_{i=1}^{n} P_i(\nearrow) P_i(\rightarrow) \tag{2}$$

Note that $P_i(\nearrow)$ is just a shorthand for $P(U_i = 1 + U_{i-1})$, the probability of seeing a new word on step i, and $P_i(\rightarrow) = 1 - P_i(\nearrow)$.

2.2 Sampling from a fixed distribution

From equation (1), in order to compute expected vocabulary size $E[U_n]$, we need accurate estimates for $P_i(\nearrow)$, the probability of a new word arriving at step i. We assume that words are randomly sampled from some possibly infinite set $\mathcal{W}$. Suppose W_i, the ith observed word, is w. The probability that w is a new word is just the probability that we have not seen w before the ith word:

$$P_i(\nearrow | W_i = w) = P(\forall_{k<i} W_k \neq w)$$

Now it is easy to express $P_i(\nearrow)$ as the expectation over all possible words that could come at step i:

$$P_i(\nearrow) = \sum_{w \in \mathcal{W}} P(W_i = w) P(\forall_{k<i} W_k \neq w)$$

If we further assume that the words W_i are sampled independently of each other from the common distribution $P(W_i = w) = p_w$, we arrive at the following formulation:

$$P_i(\nearrow) = \sum_{w \in \mathcal{W}} p_w (1 - p_w)^{i-1} \tag{3}$$

Note that equation (3) has an intuitive interpretation: probability of seeing a new word at step i is the expectation of not seeing some word $(i-1)$ times in a row. Now we can apply equation (3) to equation (1) and go through the following derivation:

$$E[U_n] = \sum_{i=1}^{n} \sum_{w \in \mathcal{W}} p_w (1-p_w)^{i-1} = \sum_{w \in \mathcal{W}} p_w \frac{1 - (1-p_w)^n}{1 - (1-p_w)}$$

The last step was obtained by changing the order of summations and using an algebraic closed form for $\sum_{i=0}^{n-1} b^i$. Now we can write down a simple formulation for the expected vocabulary size, given the relative frequencies with which we expect to see different words w:

$$E[U_n] = \sum_{w \in \mathcal{W}} (1 - (1 - p_w)^n) \tag{4}$$

2.3 Estimation of word probabilities

The final step in our model is the estimation of probability with which we expect to see a given word w. In the field of language processing, there is a large body of work on modeling word frequencies. The most famous result is Zipf's law [7], which states that the frequency of a given word w multiplied by its rank r is a constant for all words. The rank is obtained by sorting all the words in decreasing order of frequency. An extension of Zipf's law, suggested by Mandelbrot [4] gives:

$$p_w \approx \left(\frac{c}{r + a} \right)^b \tag{5}$$

While Zipf's law describes word frequencies in general, it is rather a rather poor fit to any given collection. Figure 2 shows on a logarithmic scale the empirical distribution of word probabilities in TREC volumes 1 and 2 (circles). A dotted line is a fit of generalized Zipf's equation (5) to the data, where a was set to 0. It is obvious that Zipf's law provides a reasonable fit for the majority of the words, but is really poor in matching the most common words. A solid line, which gives a better overall fit is given by fitting the following distribution to the data:

$$p_w \approx 1 - \left(\frac{r^2}{r^2 + a} \right)^b \tag{6}$$

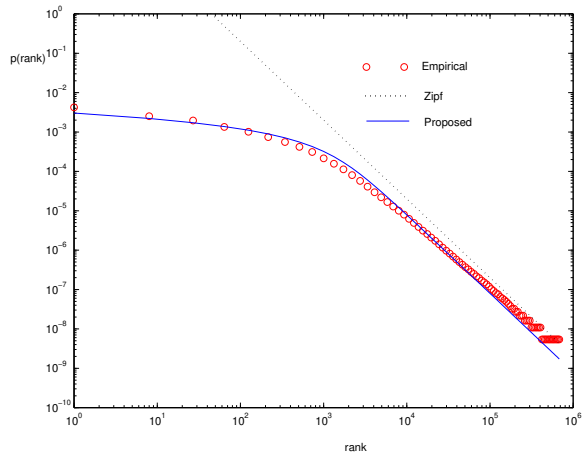

Figure 2: Word probabilities as a function of word rank. Empirical word probabilities in TREC volumes 1 and 2 are shown with circles. All the words are stemmed, with 400 *stopwords* removed. Zipf formula (5) is a good fit for the tail of the distribution, but fails to capture high-frequency terms. Equation (6) provides a good overall fit.

Here r is the rank of w, while a and b are the parameters that jointly control the curvature of the distribution and normalize p_w so that $\sum_w p_w = 1$. Another convenient property of equation (6) is that $\sum_w p_w$ converges even if number of words is infinite. We will use this property in assuming that we are sampling from an infinitely large set of words $\mathcal{W}$.

Aside from providing a better empirical fit to the word frequencies in our dataset, equation (6) allows an elegant closed-form solution for the expected size of vocabulary. If we substitute our estimate for p_w into equation (4), we get the following:

$$E[U_n] = \sum_{w \in \mathcal{W}} (1-(1-p_w)^n) \approx \int_1^{\infty} 1-\left(\frac{r^2}{r^2+a}\right)^{bn} dr$$

While there is no closed form solution for the summation above, it turns out that the integral above is the exact expansion of the Gauss hypergeometric function ${}_2F_1$ [1]:

$$E[U_n] \approx {}_2F_1\left(-\frac{1}{2}, bn, \frac{1}{2}, -a\right) \qquad (7)$$

The main result of this section is the following: if we assume that words are independently sampled from equation (6), the expected number of unique words follows a hypergeometric function ${}_2F_1$.

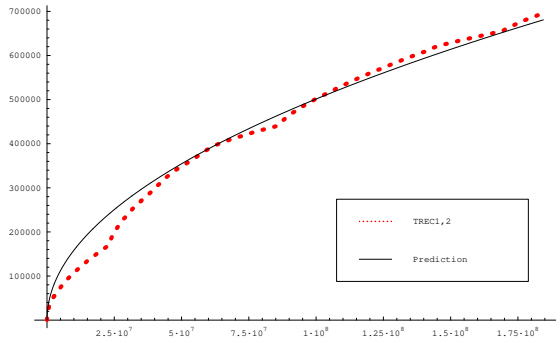

Figure 3: Vocabulary size as a function of number of words in a collection. Dotted line shows empirical vocabulary growth in TREC volumes 1 and 2. Solid line shows our estimate of vocabulary size.

3 Experimental Results

We performed a number of experiments to check the ability of our model to match the empirical growth of vocabulary. Figure 3 shows the growth on the data in TREC volumes 1 and 2 along with the expected value from our model. Note that we measure vocabulary growth of stemmed words.

The model provides a good overall fit to the growth. The parameter values a and b for equation (7) were selected to match the empirical word distribution in Figure 3. As expected, the model does not capture the artificial "dips" in the growth. The dips happen because documents in TREC are ordered not randomly, but according to source. The dips occur precisely when we shift from one source to another (e.g. from AP Newswire to the Federal Register). Figure 4 shows model predictions on the AP'88 subset of our dataset. The overall fit appears much better, because AP'88 does not exhibit abrupt shifts in sources of documents.

3.1 Confidence Interval

Our model allows us not only to predict the expected value of vocabulary size, but also to get a confidence interval on the deviations of actual vocabulary size from the expected value. From equation (1), we can estimate the upper bound on the standard deviation of U_n:

$$\sigma(U_n) = \sqrt{VU_n} \leq \sqrt{E[U_n]}$$

Since we don't know the distribution of U_n, we cannot apply parametric methods to compute the confidence interval around $E[U_n]$. However, we can use Chebyshev's theorem [6] to assert that:

$$P(|U_n - E[U_n]| > k\sqrt{E[U_n]}) \leq \frac{1}{k^2} \qquad (8)$$

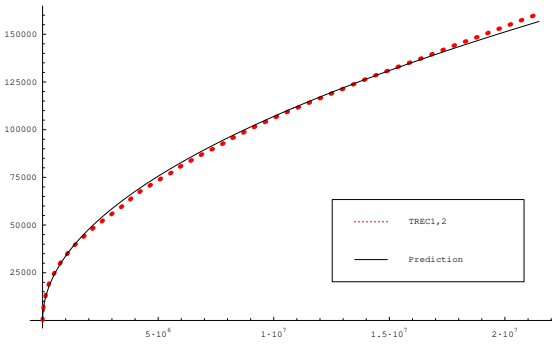

Figure 4: Vocabulary size as a function of number of words in a collection. Dotted line shows empirical vocabulary growth in AP'88 dataset. Solid line shows our estimate of vocabulary size.

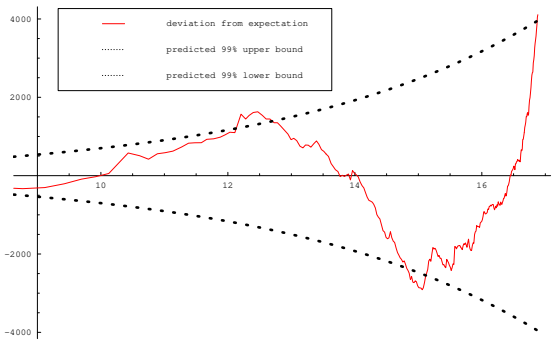

Figure 5: Deviation of observed vocabulary size (AP'88) from our estimate. Dotted lines show a 99% confidence interval, computed from equation (8). x-axis represents total number of words in the collection (note the logarithmic scale).

Equation (8) allows us to construct a 99% confidence interval around the expected vocabulary size. The confidence interval is shown in Figure 5, along with the actual deviation of vocabulary size from our prediction. The logarithmic scale of the total size of the dataset highlights the fact that confidence interval holds when n is small as well. Note that apparent exponential growth of the deviations is an artifact of logarithmic scale; actual deviations grow as the square root of the total vocabulary size (equation (8)).

4 Summary

We presented a mathematical relation between the total size of the textual dataset and the number of distinct words in that text. Our model allows us to compute the expected size of the vocabulary and to give bounds on how much empirical data will vary around the expected value. Results in Sections 2.1 and 2.2 are not limited to vocabulary growth and are applicable to a wide class of sampling tasks. Our model is able to accurately capture vocabulary growth in homogeneous datasets (e.g. AP'88). For heterogeneous sets (e.g. TREC volumes 1 and 2), the model is accurate except when abrupt shifts of source occur in the collection. The main result of this work is that under reasonable assumptions the vocabulary growth follows Gauss' hypergeometric function ${}_2F_1$.

5 Acknowledgments

The author would like to thank Bruce Croft, S. Ravela and R. Manmatha for helpful discussions of this work. This material is based on work supported in part by the Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

[1] Irene Stegun Editors: Milton Abramowitz. *Handbook of Mathematical Functions*, pages 556–560. Dover Publications, Inc., New York, 1972.

[2] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*, pages 206–208. Academic Press, Inc., New York, 1978.

[3] Bruce Croft et.al. Jamie Callan. *Advances in Information Retrieval*, pages 138–143. Kluwer Academic Publishers, New York, 2000.

[4] Benoit Mandelbrot. *Structure Formelle des Textes et Communication*. Word10:1-27, 1954.

[5] Jay Ponte. *A Language Modeling Approach to Information Retrieval*. PhD thesis, Dept. of Computer Science, University of Massachusetts, Amherst, 1998.

[6] Raymond H. Myers Ronald E. Walpole. *Probability and Statistics for Engineers and Scientists*, pages 108–109. MacMillan Publishing Company, New York, 1989.

[7] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison Wesley, Cambridge, MA, 1949.