Perplexity on Reduced Corpora

Hayato Kobayashi*
Yahoo Japan Corporation
9-7-1 Akasaka, Minato-ku, Tokyo 107-6211, Japan
hakobaya@yahoo-corp.jp

Abstract

This paper studies the idea of removing low-frequency words from a corpus, which is a common practice to reduce computational costs, from a theoretical standpoint. Based on the assumption that a corpus follows Zipf's law, we derive trade-off formulae of the perplexity of k-gram models and topic models with respect to the size of the reduced vocabulary. In addition, we show an approximate behavior of each formula under certain conditions. We verify the correctness of our theory on synthetic corpora and examine the gap between theory and practice on real corpora.

1 Introduction

Removing low-frequency words from a corpus (often called *cutoff*) is a common practice to save on the computational costs involved in learning language models and topic models. In the case of language models, we often have to remove low-frequency words because of a lack of computational resources, since the feature space of k-grams tends to be so large that we sometimes need cutoffs even in a distributed environment (Brants et al., 2007). In the case of topic models, the intuition is that low-frequency words do not make a large contribution to the statistics of the models. Actually, when we try to roughly analyze a corpus with topic models, a reduced corpus is enough for the purpose (Steyvers and Griffiths, 2007).

A natural question arises: How many low-frequency words can we remove while maintaining sufficient performance? Or more generally, by how much can we reduce a corpus/model using a certain strategy and still keep a sufficient level of performance? There have been many studies addressing the question as it pertains to different strategies (Stolcke, 1998; Buchsbaum et al., 1998; Goodman and Gao, 2000; Gao and Zhang, 2002; Ha et al., 2006; Hirsimaki, 2007; Church et al., 2007). Each of these studies experimentally discusses trade-off relationships between the size of the reduced corpus/model and its performance measured by perplexity, word error rate, and other factors. To our knowledge, however, there is no theoretical study on the question and no evidence for such a trade-off relationship, especially for topic models.

In this paper, we first address the question from a theoretical standpoint. We focus on the cutoff strategy for reducing a corpus, since a cutoff is simple but powerful method that is worth studying; as reported in (Goodman and Gao, 2000; Gao and Zhang, 2002), a cutoff is competitive with sophisticated strategies such as entropy pruning. As the basis of our theory, we assume Zipf's law (Zipf, 1935), which is an empirical rule representing a long-tail property of words in a corpus. Our approach is essentially the same as those in physics, in the sense of constructing a theory while believing experimentally observed results. For example, we can derive the distance to the landing point of a ball thrown up in the air with initial speed v_0 and angle θ as ${v_0}^2 \sin(2\theta)/g$ by believing in the experimentally observed gravity acceleration g. In a similar fashion, we will try to clarify the trade-off relationship by believing Zipf's law.

The rest of the paper is organized as follows. In Section 2, we define the notation and briefly explain Zipf's law and perplexity. In Section 3, we theoretically derive the trade-off formulae of the cutoff for unigram models, k-gram models, and topic models, each of which represents its perplexity with respect to a reduced vocabulary, under the assumption that the corpus follows Zipf's law. In addition, we show an approximate behavior of each formula under certain conditions. In

*This work was mainly carried out while the author was with Toshiba Corporation.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 797–806,
Baltimore, Maryland, USA, June 23-25 2014.

Section 4, we verify the correctness of our theory on synthetic corpora and examine the gap between theory and practice on several real corpora. Section 5 concludes the paper.

2 Preliminaries

Let us consider a corpus $\mathbf{w} := w_1 \cdots w_N$ of corpus size N and vocabulary size W. We use an abridged notation $\{\mathbf{w}\} := \{w \in \mathbf{w}\}$ to represent the vocabulary of $\mathbf{w}$. Clearly, $N = |\mathbf{w}|$ and $W = |\{\mathbf{w}\}|$ hold. When $\mathbf{w}$ has additional notations, N and W inherit them. For example, we will use N' as the size of $\mathbf{w}'$ without its definition.

2.1 Power law and Zipf's law

A power law is a mathematical relationship between two quantities x and y, where y is proportional to the c-th power of x, i.e., $y \propto x^c$, and c is a real number. Zipf's law (Zipf, 1935) is a power law discovered on real corpora, wherein for any word $w \in \mathbf{w}$ in a corpus $\mathbf{w}$, its frequency (or word count) $f(w)$ is inversely proportional to its frequency ranking $r(w)$, i.e.,

$$f(w) = \frac{C}{r(w)}.$$

Here, $f(w) := |\{w' \in \mathbf{w} \mid w' = w\}|$, and $r(w) := |\{w' \in \mathbf{w} \mid f(w') \geq f(w)\}|$. From the definition, the constant C is the maximum frequency in the corpus. Taking the natural logarithms $\ln(\cdot)$ of both sides of the above equation, we find that its plot becomes linear on a log-log graph of $r(w)$ and $f(w)$. In fact, the result based on a statistical test in (Clauset et al., 2009) reports that the frequencies of words in a corpus completely follow a power law, whereas many datasets with long-tail properties, such as networks, actually do not follow power laws.

2.2 Perplexity

Perplexity is a widely used evaluation measure of k-gram models and topic models. Let p be a predictive distribution over words, which was learned from a training corpus $\mathbf{w}$ based on a certain model. Formally, perplexity *PP* is defined as the geometric mean of the inverse of the per-word likelihood on the held-out test corpus $\mathbf{w}_\tau$, i.e.,

$$PP := \left(\prod_{w \in \mathbf{w}_\tau} \frac{1}{p(w)} \right)^{\frac{1}{N_\tau}}.$$

Intuitively, *PP* means how many possibilities one has for estimating the next word in a test corpus. According to the definition, a lower perplexity means better generalization performance of p. Another well-known evaluation measure is cross-entropy. Since cross-entropy is easily calculated as $\log_2 PP$, we can apply many of the results of this paper to cross-entropy.

3 Perplexity on Reduced Corpora

Now let us consider what a cutoff is. In our study, we simply define a corpus that has been reduced by removing low-frequency words from the original corpus with a certain threshold. Formally, we say $\mathbf{w}'$ is a *corpus reduced from the original corpus* $\mathbf{w}$, if $\mathbf{w}'$ is the longest subsequence of $\mathbf{w}$ such that $\max_{w' \in \mathbf{w}'} r(w') = W'$. Note that a subsequence can include gaps in contrast to a substring. For example, supposing we have a corpus $\mathbf{w} = abcaba$ with a vocabulary $\{\mathbf{w}\} = \{a, b, c\}$, $\mathbf{w}'_1 = ababa$ is a reduced corpus, while $\mathbf{w}'_2 = aba$ and $\mathbf{w}'_3 = acaa$ are not.

After learning a distribution p' from a reduced corpus $\mathbf{w}'$, we need to infer the distribution p learned from the original corpus $\mathbf{w}$. Here, we use *constant restoring* (defined below), which assumes the frequencies of the reduced low-frequency words are a constant.

Definition 1 (Constant Restoring)**.** *Given a positive constant λ, a distribution p' over a reduced corpus $\mathbf{w}'$, and a corpus $\mathbf{w}$, we say that $\hat{p}$ is a λ-*restored distribution of p' from $\mathbf{w}'$ to $\mathbf{w}$*, if $\sum_{w \in \{\mathbf{w}\}} \hat{p}(w) = 1$, and for any $w \in \mathbf{w}$,*

$$\hat{p}(w) \propto \begin{cases} p'(w) & (w \in \mathbf{w}') \\ \lambda & (w \notin \mathbf{w}'). \end{cases}$$

Constant restoring is similar to the additive smoothing defined by $\hat{p}(w) \propto p'(w) + \lambda$, which is used to solve the zero-frequency problem of language models (Chen and Goodman, 1996). The only difference is the addition of a constant λ only to zero-frequency words. We think constant restoring is theoretically natural in our setting, since we can derive the above equation by letting each frequency of reduced words be $\lambda N'$ and defining a restored frequency function as follows:

$$\hat{f}(w) = \begin{cases} f(w) & (w \in \mathbf{w}') \\ \lambda N' & (w \notin \mathbf{w}'). \end{cases}$$

Informally, constant restoring involves padding the vocabulary, while additive smoothing involves padding the corpus. Smoothing should be carried out after restoring.

3.1 Perplexity of Unigram Models

Let us consider the perplexity of a unigram model learned from a reduced corpus. In unigram models, a predictive distribution p' on a reduced corpus $\mathbf{w}'$ can be simply calculated as $p'(w') = f(w')/N'$. We shall start with an analysis of training-set perplexity, since we can derive an exact formula for it, which will give us a sufficient idea for making an approximate analysis of test-set perplexity.

Let $\hat{PP}_1 := \left(\prod_{w\in\mathbf{w}} \frac{1}{\hat{p}(w)}\right)^{\frac{1}{N}}$ be the perplexity of a λ-restored distribution $\hat{p}$ on a unigram model. The next lemma gives the optimal restoring constant λ^* minimizing $\hat{PP}_1$.

Lemma 2. *For any λ-restored distribution $\hat{p}$ of a distribution p' from a reduced corpus $\mathbf{w}'$ to the original corpus $\mathbf{w}$, its perplexity is minimized by*

$$\lambda^* = \frac{N - N'}{(W - W')N'}.$$

Proof. Let $\mathbf{w}_\mathcal{R}$ be the longest subsequence such that $\min_{w'\in\mathbf{w}'} r(w') = W' + 1$. Since $\mathbf{w}_\mathcal{R}$ is the remainder of $\mathbf{w}'$, $N_\mathcal{R} = N - N'$ and $W_\mathcal{R} = W - W'$ hold. After substituting the normalized form of $\hat{p}$ of Definition 1 into $\hat{PP}_1$, we have

$$\begin{aligned}\hat{PP}_1 &= \left(\prod_{w'\in\mathbf{w}'} \frac{1}{\hat{p}(w')} \prod_{w_\mathcal{R}\in\mathbf{w}_\mathcal{R}} \frac{1}{\hat{p}(w_\mathcal{R})}\right)^{\frac{1}{N}} \\ &= \left(\prod_{w'\in\mathbf{w}'} \frac{1 + W_\mathcal{R}\lambda}{p'(w')} \prod_{w_\mathcal{R}\in\mathbf{w}_\mathcal{R}} \frac{1 + W_\mathcal{R}\lambda}{\lambda}\right)^{\frac{1}{N}} \\ &= \frac{1 + W_\mathcal{R}\lambda}{\lambda^{\frac{N_\mathcal{R}}{N}}} \left(\prod_{w'\in\mathbf{w}'} \frac{1}{p'(w')}\right)^{\frac{1}{N}}.\end{aligned}$$

We obtain the optimal smoothing factor λ^* when $\frac{\partial}{\partial\lambda}\hat{PP}_1 \propto \frac{\partial}{\partial\lambda}(1 + W_\mathcal{R}\lambda)/\lambda^{\frac{N_\mathcal{R}}{N}} = 0$. □

By using a similar argument to the one in the above lemma, we can obtain the optimal constant of additive smoothing as $\lambda^* \approx \frac{N-N'}{WN'}$, when N is sufficiently large.

The next theorem gives the exact formula of the training-set perplexity of a unigram model learned from a reduced corpus.

Theorem 3. *For any distribution p' on a unigram model learned from a corpus $\mathbf{w}'$ reduced from the original corpus $\mathbf{w}$ following Zipf's law, the perplexity $\hat{PP}_1$ of the λ^*-restored distribution $\hat{p}$ of p' from $\mathbf{w}'$ to $\mathbf{w}$ is calculated by*

$$\begin{aligned}\hat{PP}_1(W') =& H(W) \exp\left(\frac{B(W')}{H(W)}\right) \\ & \left(\frac{W - W'}{H(W) - H(W')}\right)^{1 - \frac{H(W')}{H(W)}},\end{aligned}$$

where $H(X) := \sum_{x=1}^{X} \frac{1}{x}$ and $B(X) := \sum_{x=1}^{X} \frac{\ln x}{x}$.

Proof. We expand the first part of $\hat{PP}_1$ in the proof of Lemma 2 using λ^* as follows:

$$\begin{aligned}\frac{1 + W_\mathcal{R}\lambda^*}{{\lambda^*}^{\frac{N_\mathcal{R}}{N}}} &= \left(1 + \frac{N_\mathcal{R}}{N'}\right)\left(\frac{W_\mathcal{R}N'}{N_\mathcal{R}}\right)^{\frac{N_\mathcal{R}}{N}} \\ &= \left(\frac{N}{N'}\right)\left(\frac{(W - W')N'}{N - N'}\right)^{1 - \frac{N'}{N}}.\end{aligned}$$

The second part of $\hat{PP}_1$ is as follows:

$$\begin{aligned}\left(\prod_{w'\in\mathbf{w}'} \frac{1}{p'(w')}\right)^{\frac{1}{N}} &= \prod_{w'\in\{\mathbf{w}'\}} \left(\frac{1}{p'(w')}\right)^{\frac{f(w')}{N}} \\ &= \prod_{r=1}^{W'} \left(\frac{rN'}{C}\right)^{\frac{C}{rN}} \\ &= \prod_{r=1}^{W'} \left(\frac{N'}{C}\right)^{\frac{C}{rN}} \prod_{r=1}^{W'} r^{\frac{C}{rN}} \\ &= \left(\frac{N'}{C}\right)^{\frac{N'}{N}} \exp\left(\frac{C}{N}\sum_{r=1}^{W'} \frac{\ln r}{r}\right).\end{aligned}$$

We obtain the objective formula by putting the above two formulae together with $N = CH(W)$ and $N' = CH(W')$, which are derived from Zipf's law. □

The functions $H(X)$ and $B(X)$ are the X-th partial sum of the harmonic series and Bertrand series (special form), respectively. An approximation by definite integrals yields $H(X) \approx \ln X + \gamma$, where γ is the Euler-Mascheroni constant, and $B(X) \approx \frac{1}{2}\ln^2 X$. We may omit γ from the approximate analysis.

Now let us consider an approximate form of $\hat{PP}_1(W')$ in Theorem 3. For further discussion,

we define the last part of $\hat{PP}_1(W')$ as follows:

$$F(W, W') := \left(\frac{W - W'}{H(W) - H(W')}\right)^{1-\frac{H(W')}{H(W)}}.$$

Since $W' = \delta W$ holds for an appropriate ratio δ, we have

$$\begin{aligned} F(W, \delta W) &= \left(\frac{W - \delta W}{H(W) - H(\delta W)}\right)^{1-\frac{H(\delta W)}{H(W)}} \\ &\approx \left(\frac{W - \delta W}{\ln W - \ln(\delta W)}\right)^{1-\frac{\ln(\delta W)}{\ln W}} \\ &= \left(\frac{W(1-\delta)}{-\ln \delta}\right)^{\frac{-\ln \delta}{\ln W}} \\ &\to \frac{1}{\delta} \qquad (W \to \infty). \end{aligned}$$

Therefore, when W is sufficiently large, we can use $F(W, W') \approx \frac{W}{W'}$, since $F(W, \delta W) \approx \frac{1}{\delta}$ holds for any ratio $\delta : 0 < \delta < 1$. Using this fact, we obtain an approximate formula $\tilde{PP}_1$ of $\hat{PP}_1$ as follows:

$$\begin{aligned} \tilde{PP}_1(W') &= \ln W \exp\left(\frac{\ln^2 W'}{2 \ln W}\right) \frac{W}{W'} \\ &= \sqrt{W} \ln W \exp \frac{(\ln W' - \ln W)^2}{2 \ln W}. \end{aligned}$$

The complexity of $\tilde{PP}_1$ is quasi-polynomial, i.e., $\tilde{PP}_1(W') = O(W'^{\ln W'})$, which behaves as a quadratic function on a log-log graph. Since $\tilde{PP}_1(W')$ is convex, i.e., $\frac{\partial^2}{\partial W'^2}\tilde{PP}_1(W') > 0$, and its gradient $\frac{\partial}{\partial W'}\tilde{PP}_1(W')$ is zero when $W' = W$, we infer that low-frequency words may not largely contribute to the statistics.

Considering the special case of $W' = W$, we obtain the perplexity PP_1 of the unigram model learned from the original corpus $\mathbf{w}$ as

$$PP_1 = H(W) \exp\left(\frac{B(W)}{H(W)}\right) \approx \sqrt{W} \ln W.$$

Interestingly, PP_1 is approximately expressed as a simple elementary function of vocabulary size W. This suggests that models learned from corpora with the same vocabulary size theoretically have the same perplexity.

For the test-set perplexity, we assume that both the training corpus $\mathbf{w}$ and test corpus $\mathbf{w}_\tau$ are generated from the same distribution based on Zipf's law. This assumption is natural, considering the situation of an in-domain test or cross validation test. Let $\mathbf{w}_\tau{}'$ be the longest subsequence of $\mathbf{w}_\tau$ such that for any $w \in \mathbf{w}_\tau{}'$, $w \in \mathbf{w}'$ holds. Formally, we assume $p'(w) \approx p_\tau{}'(w)$ for any $w \in \mathbf{w}'_\tau$ when $W_\tau > W'$, where $p_\tau{}'$ is the true distribution over $\mathbf{w}_\tau{}'$. Using similar arguments to those of Lemma 2 and Theorem 3 for $\mathbf{w}_\tau$, we obtain an approximation formula for the test-set perplexity, where we simply substitute W and W' in the exact formula for the training-set perplexity with W_τ and $W_\tau{}'$, respectively. For simplicity, we will only consider training-set perplexity from now on, since we can make a similar argument for the test-set perplexity in the later analysis.

3.2 Perplexity of k-gram Models

Here, we will consider the perplexity of a k-gram model learned from a reduced corpus as a standard extension of a unigram model. Our theory only assumes that the corpus is generated on the basis of Zipf's law. Thus, we can use a simple model where k-grams are calculated from a random word sequence based on Zipf's law. This model seems to be stupid, since we can easily notice that the bigram "is is" is quite frequent, and the two bigrams "is a" and "a is" have the same frequency. However, the experiments described later uncovered the fact that the model can roughly capture the behavior of real corpora.

The frequency f_k of k-gram word $w_k \in \mathbf{w}^k$ in the model is represented by the following formula:

$$f_k(w_k) = \frac{C_k}{g_k(r_k(w_k))},$$

where C_k is the maximal frequency in k-grams, r_k is the frequency ranking of w_k over k-grams, and g_k expresses the frequency decay in k-grams. For example, the decay function g_2 of bigrams is as follows:

$$\begin{aligned} (g_2(i))_i &:= (g_2(1),\ g_2(2),\ g_2(3), \cdots) \\ &= (1 \cdot 1,\ 1 \cdot 2,\ 2 \cdot 1,\ 1 \cdot 3,\ 3 \cdot 1, \cdots) \\ &= (1,\ 2,\ 2,\ 3,\ 3,\ 4,\ 4,\ 4,\ 5,\ 5,\ 6, \cdots). \end{aligned}$$

This is an inverse of the sum of Piltz's divisor functions $d_2(n) := \sum_{i_1 \cdot i_2 = n} 1$, which represents the number of divisors of an integer n (cf. (OEIS, 2001)). In general, we formally define g_k through its inverse: $g_k^{-1}(\ell) := S_k(\ell)$, where $S_k(\ell) := \sum_{n=1}^{\ell} d_k(n)$ and $d_k(n) := \sum_{i_1 \cdot i_2 \cdots i_k = n} 1$. Since $(g_k(i))_i$ is a sorted sequence of the elements of the k-th tensor power of vector $(1, \cdots, W)$, we can calculate the maximum frequency C_k as follows.

Lemma 4. *For any corpus* $\mathbf{w}$ *following Zipf's law, the maximum frequency of* k*-grams in our model is calculated by*

$$C_k = \frac{N-(k-1)D}{(H(W))^k},$$

where D *denotes the number of documents in* $\mathbf{w}$.

Proof. We use $\sum_{w_k} f_k(w_k) = C_k(\sum_w 1/r(w))^k$. □

The sum $S_k(\ell)$ of Piltz's divisor functions can be approximated by $\ell P_k(\ln \ell)$, where $P_k(x)$ is a polynomial of degree $k-1$ with respect to x, and the main term of $\ell P_k(\ln \ell)$ is given by the following residue $\mathrm{Res}_{s=1}\frac{\zeta^k(s)x^s}{s}$, where $\zeta(s)$ is the Riemann zeta function (Li, 2005). Using this fact, we obtain an approximation $\ln\left(g_k^{-1}(\ell)\right) \approx \ln \ell + O(\ln(\ln \ell)) \approx \ln \ell$, when ℓ is sufficiently large. Thus, when the corpus is sufficiently large, we can see that the behavior of f_k is roughly linear on a log-log graph, i.e., $f_k(w_k) \propto r_k(w_k)^{-1}$, since if $g_k^{-1}(\ell) \propto \ell^c$ holds, then $f_k(r) \propto (g_k(r))^{-1} \propto r^{-\frac{1}{c}}$ holds.

Unfortunately, however, most corpora in the real world are not so large that the above-mentioned relation holds. Actually, Ha et al. (Ha et al., 2002; Ha et al., 2006) experimentally found that although a k-gram corpus roughly follows a power law even when $k > 1$, its exponent is smaller than 1 (for Zipf's law). They pointed out that the exponent of bigrams is about 0.66, and that of 5-grams is about 0.59 in the Wall Street Journal corpus (WSJ87). Believing their claim that there exists a constant π_k such that $f_k(w_k) \propto r_k(w_k)^{-\pi_k}$, we estimated the exponent of k-grams in an actual situation in the form of the following lemma.

Lemma 5. *Assuming that* $f_k(w_k) \propto r_k(w_k)^{-\pi_k}$ *holds for any* k*-gram word* $w_k \in \mathbf{w}^k$ *in a corpus* $\mathbf{w}$ *following Zipf's law, the optimal exponent in our model based on the least squares criterion is calculated by*

$$\pi_k = \frac{\ln W}{(k-1)\ln(\ln W) + \ln W}.$$

Proof. We find the optimal exponent π_k by minimizing the sum of squared errors between the gradients of $g_k^{-1}(r)$ and $r^{\frac{1}{\pi_k}}$ on a log-log graph:

$$\int \left\{ \frac{\partial}{\partial y}\left(y + \ln P_k(y)\right) - \frac{\partial}{\partial y}\left(\frac{1}{\pi_k}y\right)\right\}^2 dy,$$

where $y = \ln r$. □

In the case of unigrams ($k = 1$), the formula exactly represents Zipf's law. In the case of k-grams ($k > 1$), we found that the formula approaches Zipf's law when W approaches infinity, i.e., $\lim_{W\to\infty} \pi_k = 1$.

Let us consider the perplexity of a k-gram model learned from a reduced corpus. We immediately obtain the following corollary using Lemma 5.

Corollary 6. *For any distribution* p' *on a* k*-gram model learned from a corpus* $\mathbf{w}'$ *reduced from the original corpus* $\mathbf{w}$ *following Zipf's law, assuming that* $f_k(w_k) \propto r_k(w_k)^{-\pi_k}$ *holds for any* k*-gram word* $w_k \in \mathbf{w}^k$ *and the optimal exponent* π_k *in Lemma 5, the perplexity* $\hat{PP}_k$ *of the* λ^**-restored distribution* $\hat{p}$ *of* p' *from* $\mathbf{w}'$ *to* $\mathbf{w}$ *is calculated by*

$$\hat{PP}_k(W') = H_{\pi_k}(W)\exp\left(\frac{B_{\pi_k}(W')}{H_{\pi_k}(W)}\right) \left(\frac{W-W'}{H_{\pi_k}(W)-H_{\pi_k}(W')}\right)^{1-\frac{H_{\pi_k}(W')}{H_{\pi_k}(W)}},$$

where $H_a(X) := \sum_{x=1}^{X}\frac{1}{x^a}$ *and* $B_a(X) := \sum_{x=1}^{X}\frac{a\ln x}{x^a}$.

$H_a(X)$ is the X-th partial sum of the P-series or hyper-harmonic series, which is a generalization of the harmonic series $H(X)$. $B_a(X)$ is the X-th partial sum of the Bertrand series (another special form of $B(X)$). When $0 < a < 1$, we can easily calculate $\hat{PP}_k(W')$ by using the following approximations:

$$H_a(X) \approx \frac{(X+1)^{1-a}-1}{1-a}$$
$$B_a(X) \approx \frac{a}{1-a}(X+1)^{1-a}\ln(X+1) - \frac{a}{(1-a)^2}(X+1)^{1-a} + \frac{a}{(1-a)^2}.$$

By putting the approximations of $H_a(X)$ and $B_a(X)$ into the formula of Corollary 6, we obtain an approximation $\hat{PP}_k(W') \approx O({W'}^{{W'}^{1-\pi_k}})$. This implies that $\hat{PP}_k(W')$ is approximately linear on a log-log graph, when π_k is close to 1, i.e., k is relatively small and W is sufficiently large. Note that we must use the approximation of $H(X)$, not $H_a(X)$, when $a = 1$.

The fact that the frequency of k-grams follows a power law leads us to an additional convenient

property, since the process of generating a corpus in our theory can be treated as a variant of the coupon collector's problem. In this problem, we consider how many trials are needed for collecting all coupons whose occurrence probabilities follow some stable distribution. According to a well-known result about power law distributions (Boneh and Papanicolaou, 1996), we need a corpus of size $\frac{kW^k}{1-\pi_k}\ln W$ when $\pi_k < 1$, and $W\ln^2 W$ when $\pi_k = 1$ for collecting all of the k-grams, the number of which is W^k. Using results in (Atsonios et al., 2011), we can easily obtain a lower and upper bound of the actual vocabulary size $\tilde{W}_k$ of k-grams from the corpus size N and vocabulary size W as

$$\tilde{W}_k \geq (\pi_k + 1)\left(1 - e^{-\frac{(1-\pi_k)N}{W^k-1} - \ln\frac{W^k-1}{W^k}}\right)$$

$$\tilde{W}_k \leq \frac{\pi_k}{\pi_k - 1}\left(\frac{N}{H_{\pi_k}(W^k)}\right)^{\frac{1}{\pi_k}} - \frac{NW^{1-\pi_k}}{(\pi_k - 1)H_{\pi_k}(W^k)}.$$

This means that we can determine the rough sparseness of k-grams and adjust some of the parameters such as the gram size k in learning statistical language models.

3.3 Perplexity of Topic Models

In this section, we consider the perplexity of the widely used topic model, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), by using the notation given in (Griffiths and Steyvers, 2004). LDA is a probabilistic language model that generates a corpus as a mixture of hidden topics, and it allows us to infer two parameters: the document-topic distribution θ that represents the mixture rate of topics in each document, and the topic-word distribution ϕ that represents the occurrence rate of words in each topic. For a given corpus $\mathbf{w}$, the model is defined as

$$\begin{aligned} \theta_{d_i} &\sim \textit{Dirichlet}(\alpha) \\ z_i|\theta_{d_i} &\sim \textit{Multi}(\theta_{d_i}) \\ \phi_{z_i} &\sim \textit{Dirichlet}(\beta) \\ w_i|z_i, \phi_{z_i} &\sim \textit{Multi}(\phi_{z_i}), \end{aligned}$$

where d_i and z_i are respectively the document that includes the i-th word w_i and the hidden topic that is assigned to w_i. In the case of inference by Gibbs sampling presented in (Griffiths and Steyvers, 2004), we can sample a "good" topic assignment z_i for each word w_i with high probability. Using the assignments $\mathbf{z}$, we obtain the posterior distributions of two parameters as $\hat{\theta}_d(z) \propto n_z^{(d)} + \alpha$ and $\hat{\phi}_z(w) \propto n_z^{(w)} + \beta$, where $n_z^{(d)}$ and $n_z^{(w)}$ respectively represent the number of times assigning topic z in document d and the number of times topic z is assigned to word w.

Since an exact analysis is very hard, we will place rough assumptions on $\hat{\phi}$ and $\hat{\theta}$ to reduce the complexity. The assumption placed on $\hat{\phi}$ is that the word distribution $\hat{\phi}_z$ of each topic z follows Zipf's law. We think this is acceptable since we can regard each topic as a corpus that follows Zipf's law. Since $\hat{\phi}_z$ is normalized for each topic, we can assume that for any two topics, z and z', and any two words, w and w', $\hat{\phi}_z(w) \approx \hat{\phi}_{z'}(w')$ holds if $r_z(w) = r_{z'}(w')$, where $r_z(w)$ is the frequency ranking of w with respect to $n_z^{(w)}$. Note that the above assumption pertains to a posterior, and we do not discuss the fact that a Pitman-Yor process prior is better suited for a power law (Goldwater et al., 2011).

The assumption placed on $\hat{\phi}$ may not be reasonable in the case of $\hat{\theta}$, because we can easily think of a document with only one topic, and we usually use a small number T of topics for LDA, e.g., $T = 20$. Thus, we consider two extreme cases. One is where each document evenly has all topics, and the other is where each document only has one topic. Although these two cases might be unrealistic, the actual (theoretical) perplexity is expected to be between their values. We believe that analyzing such extreme cases is theoretically important, since it would be useful for bounding the computational complexity and predictive performance.

We can regard the former case as a unigram model, since the marginal predictive distribution $\sum_{z=1}^{T}\hat{\theta}_d(z)\hat{\phi}_z(w) \propto \sum_{z=1}^{T}\frac{n_z^{(w)}+\beta}{T} \propto\!\!\propto f(w)$ is independent of d; here we have used $\hat{\theta}_d(z) = 1/T$ from the assumption. In the latter case, we can obtain an exact formula for the perplexity of LDA when the topic assigned to each document follows a discrete uniform distribution, as shown in the next theorem. Note that a mixture of corpora following Zipf's law can be approximately regarded as following Zipf's law, when W is sufficiently large.

Theorem 7. *For any distribution p' on the LDA model with T topics learned from a corpus $\mathbf{w}'$ reduced from the original corpus $\mathbf{w}$ following Zipf's law, assuming that each document only has one topic which is assigned based on a discrete uniform distribution, the perplexity $\hat{PP}_{Mix}$ of the λ^*-restored distribution $\hat{p}$ of p' from $\mathbf{w}'$ to $\mathbf{w}$ is calcu-*

Table 1: Details of Reuters, 20news, Enwiki, Zipf1, and ZipfMix.

	vocab. size	corpus size	doc. size
Reuters	70,258	2,754,800	18,118
20news	192,667	4,471,151	19,997
Enwiki	409,902	16,711,226	51,231
Zipf1	69,786	2,754,800	18,118
ZipfMix	70,093	2,754,800	18,118

lated by

$$\hat{PP}_{Mix}(W') = H(W/T)\exp\left(\frac{B(W'/T)}{H(W/T)}\right)\left(\frac{W-W'}{H(W/T)-H(W'/T)}\right)^{1-\frac{H(W'/T)}{H(W/T)}}$$

Proof. We can prove this by using a similar argument to that of Theorem 3 for each topic. □

The formula of the theorem is nearly identical to the one of Theorem 3 for a $1/T$ corpus. This implies that the growth rate of the perplexity of LDA models is larger than that of unigram models, whereas the perplexity of LDA models for the original corpus is smaller than that of unigram models. In fact, a similar argument to the one in the approximate analysis in Section 3.1 leads to an approximate formula $\tilde{PP}_{Mix}$ of $\hat{PP}_{Mix}$ as

$$\tilde{PP}_{Mix}(W') = \sqrt{\frac{W}{T}}\ln\frac{W}{T}\exp\frac{(\ln W' - \ln W)^2}{2\ln(W/T)},$$

when W is sufficiently large. That is, $\tilde{PP}_{Mix}(W')$ also has a quadratic behavior in a log-log graph, i.e., $\tilde{PP}_{Mix}(W') = O(W'^{\ln W'})$.

4 Experiments

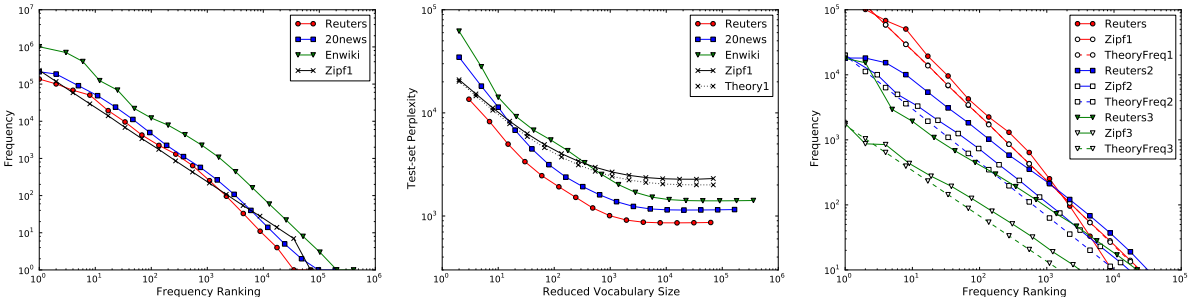
We performed experiments on three real corpora (Reuters, 20news, and Enwiki) and two synthetic corpora (Zipf1 and ZipfMix) to verify the correctness of our theory and to examine the gap between theory and practice. Reuters and 20news here denote corpora extracted from the Reuters-21578 and 20 Newsgroups data sets, respectively. Enwiki is a $1/100$ corpus of the English Wikipedia. Zipf1 is a synthetic corpus generated by Zipf's law, whose corpus is the same size as Reuters, and ZipfMix is a mixture of 20 synthetic corpora, sizes are 1/20th of Reuters. We used ZipfMix only for the experiments on topic models. Table 1 lists the details of all five corpora.

Fig. 1(a) shows the word frequency of Reuters, 20news, Enwiki, and Zipf1 versus frequency ranking on a log-log graph. In all corpora, we can regard each curve as linear with a gradient close to 1. This means that all corpora roughly follow Zipf's law. Furthermore, since the curve of Zipf1 is similar to that of Reuters, Zipf1 can be regarded as acceptable.

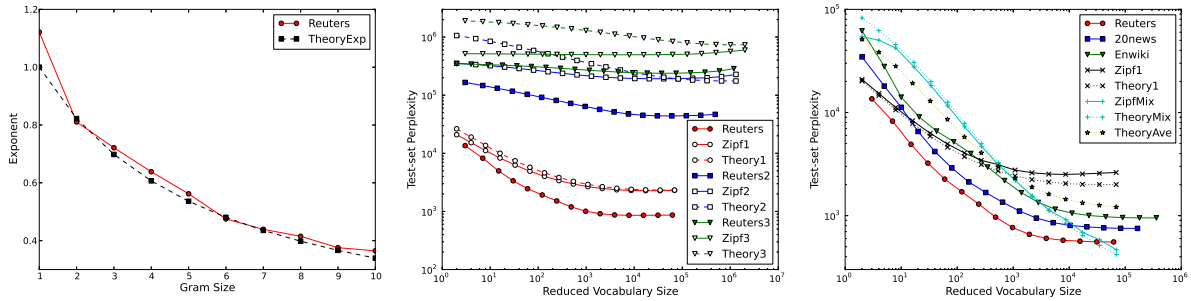
Fig. 1(b) plots the perplexity of unigram models learned from Reuters, 20news, Enwiki, and Zipf1 versus the size of reduced vocabulary on a log-log graph. Each value is the average over different test sets of five-fold cross validation. Theory1 is calculated using the formula in Theorem 3. The graph shows that the curve of Theory1 is nearly identical to that of Zipf1. Since the vocabulary size W_τ of each test set is small in this experiment, some errors appear when W' is large, i.e., $W_\tau < W'$. This clearly means that our theory is theoretically correct for an ideal corpus Zipf1. Comparing Zipf1 with Reuters, however, we find that their perplexities are quite different. The reason is that the gap between the frequencies of low-ranking (high-frequency) words is considerably large. For example, the frequency of the 1st-rank word of Reuters is $f(w) = 136,371$, while that of Zipf1 is $f(w) = 234,705$. Our theory seems to be suited for inferring the growth rate of perplexity rather than the perplexity value itself.

As for the approximate formula $\tilde{PP}_1$ of Theorem 3, we can surely regard the curve of Zipf1 as being roughly quadratic. The curves of real corpora also have a similar tendency, although their gradients are slightly steeper. This difference might have been caused by the above-mentioned errors. However, at least, we can ascertain the important fact that the results for the corpora reduced by 1/100 are not so different from those of the original corpora from the perspective of their perplexity measures.

Fig. 1(c) plots the frequency of k-grams ($k \in \{1,2,3\}$) in Reuters versus frequency ranking on a log-log graph. TheoryFreq (1-3) are calculated using C_k in Lemma 4 and π_k in Lemma 5. A comparison of TheoryFreq and Zipf verifies the correctness of our theory. However, comparing Zipf and Reuters, we see that C_k is poorly estimated when the gram size is large, whereas π_k is roughly correct. This may have happened because we did not put any assumptions on the word se-



(a) Frequency of unigrams (b) Perplexity of unigram models (c) Frequency of k-grams



(d) Exponent of a power law over k-grams (e) Perplexity of k-gram models (f) Perplexity of topic models

Figure 1: (a) Word frequency of `Reuters`, `20news`, `Enwiki`, and `Zipf1` versus frequency ranking. (b) Perplexity of unigram models learned from `Reuters`, `20news`, `Enwiki`, and `Zipf1` versus size of reduced vocabulary. `Theory1` is calculated using the formula in Theorem 3. (c) Frequency of k-grams ($k \in \{1, 2, 3\}$) in `Reuters` and `Zipf1` versus frequency ranking. The suffix digit of each label means its gram size. `TheoryFreq` (1-3) are calculated using Lemma 4 and Lemma 5. (d) Exponent of a power law over k-grams in `Reuters` versus gram size. `TheoryGrad` is calculated using π_k in Lemma 5. (e) Perplexity of k-gram models learned from `Reuters` versus size of reduced vocabulary. `Theory2` and `Theory3` are calculated using the formula in Corollary 6. (f) Perplexity of topic models learned from `Reuters`, `20news`, `Enwiki`, `Zipf1`, and `ZipfMix` versus size of reduced vocabulary. `TheoryMix` is calculated using the formula in Theorem 7.

quences in our simple model. The frequencies of high-order k-grams tend to be lower than in reality. We might need to place a hierarchical assumption on the a power law, as in done in hierarchical Pitman-Yor processes (Wood et al., 2011).

Fig. 1(d) plots the exponent of the power law over k-grams in `Reuters` versus the gram size on a normal graph. We estimated each exponent of `Reuters` by using the least-squares method. `TheoryGrad` is calculated using π_k in Lemma 5. Surprisingly, the real exponents of `Reuters` are almost the same as the theoretical estimate π_k based on our "stupid" model that does not care about the order of words. Note that we do not use any information other than the vocabulary size W and the gram size k for estimating π_k.

Fig. 1(e) plots the perplexity of k-gram models ($k \in \{1, 2, 3\}$) learned from `Reuters` versus the size of reduced vocabulary on a log-log graph. `Theory2` and `Theory3` are calculated using the formula in Corollary 6. In the case of bigrams, the perplexities of `Theory2` are almost the same as that of `Zipf2` when the size of reduced vocabulary is large. However, in the case of trigrams, the perplexities of `Theory3` are far from those of `Zipf3`. This difference may be due to the sparseness of trigrams in `Zipf3`. To verify the correctness of our theory for higher order k-gram models, we need to make assumptions that include backoff and smoothing.

Fig. 1(f) plots the perplexity of LDA models with 20 topics learned from `Reuters`, `20news`, `Enwiki`, `Zipf1`, and `ZipfMix` versus the size of reduced vocabulary on a log-log graph. We used a collapsed Gibbs sampler with 100 iterations to infer the parameters and set the hyper parameters, $\alpha = 0.1$ and $\beta = 0.1$. In evaluating the perplexity, we estimated a posterior document-topic distribu-

Table 2: Computational time and memory size for LDA learning on the original corpus, (1/10)-reduced corpus, and (1/20)-reduced corpus of `Reuters`.

corpus	time	memory	perplexity
original	4m3.80s	71,548KB	500
(1/10)	3m55.70s	46,648KB	550
(1/20)	3m42.63s	34,024KB	611

tion $\hat{\theta}_d$ by using the first half of each test document and calculated the perplexity on the second half, as is done in (Asuncion et al., 2009). Each value is the average over different test sets of five-fold cross validation. `Theory1` and `TheoryMix` are calculated using the formulae in Theorem 3 and Theorem 7, respectively. Comparing `Zipf1` with `Theory1`, and `ZipfMix` with `TheoryMix`, we find that our theory of the extreme cases discussed in Section 3.3 is theoretically correct. `TheoryAve` is the average of `Theory1` and `TheoryMix`. Comparing `Reuters` and `TheoryAve`, we see that their curves are almost the same. If theoretical perplexity $\hat{PP}$ has a similar tendency as real perplexity PP on a log-log graph, i.e., $\ln PP(W') \approx \ln \hat{PP}(W') + c$ for some constant c, we can approximate its deterioration rate as $PP(W')/PP(W) \approx \exp(\ln \hat{PP}(W') + c)/\exp(\ln \hat{PP}(W) + c) = \hat{PP}(W')/\hat{PP}(W)$. Therefore, we can use `TheoryAve` as a heuristic function for estimating the perplexity of topic models. Since we can calculate an inverse of `TheoryAve` from the bisection or Newton-Raphson method, we can maximize the reduction rate and ensure an acceptable perplexity based on a user-specified deterioration rate. According to the fact that the three real corpora with different sizes have a similar tendency, it is expected that we can use our theory for a larger corpus.

Finally, let us examine the computational costs for LDA learning. Table 2 shows computational time and memory size for LDA learning on the original corpus, (1/10)-reduced corpus, and (1/20)-reduced corpus of `Reuters`. Comparing the memory used in the learning with the original corpus and with the (1/10)-reduced corpus of `Reuters`, we find that the learning on the (1/10)-reduced corpus used 60% of the memory used by the learning on the original corpus. While the computational time decreased a little, we believe that reducing the memory size helps to reduce computational time for a larger corpus in the sense that it can relax the constraint for in-memory computing. Although we did not examine the accuracy of real tasks in this paper, there is an interesting report that the word error rate of language models follows a power law with respect to perplexity (Klakow and Peters, 2002). Thus, we conjecture that the word error rate also has a similar tendency as perplexity with respect to the reduced vocabulary size.

5 Conclusion

We studied the relationship between perplexity and vocabulary size of reduced corpora. We derived trade-off formulae for the perplexity of k-gram models and topic models with respect to the size of reduced vocabulary and showed that each formula approximately has a simple behavior on a log-log graph under certain conditions. We verified the correctness of our theory on synthetic corpora and examined the gap between theory and practice on real corpora. We found that the estimation of the perplexity growth rate is reasonable. This means that we can maximize the reduction rate, thereby ensuring an acceptable perplexity based on a user-specified deterioration rate. Furthermore, this suggests the possibility that we can theoretically derive empirical parameters, or "rules of thumb", for different NLP problems, assuming that a corpus follows Zipf's law. We believe that our theoretical estimation has the advantages of computational efficiency and scalability especially for very large corpora, although experimental estimations such as cross-validation may be more accurate.

In the future, we want to find out the cause of the gap between theory and practice and extend our theory to bridge the gap, in the same way that we can construct equations of motion with air resistance in the example of the landing point of a ball in Section 1. For example, promising research directions include using a general law such as the Zipf-Mandelbrot law (Mandelbrot, 1965), a sophisticated model that cares the order of words such as hierarchical Pitman-Yor processes (Wood et al., 2011), and smoothing/backoff methods to handle the sparseness problem.

Acknowledgments

The author would like to thank the reviewers for their helpful comments.

References

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, pages 27–34. AUAI Press.

Ioannis Atsonios, Olivier Beaumont, Nicolas Hanusse, and Yusik Kim. 2011. On power-law distributed balls in bins and its applications to view size estimation. In *Proceedings of the 22nd International Symposium on Algorithms and Computation (ISAAC 2011)*, pages 504–513. Springer-Verlag.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Shahar Boneh and Vassilis G. Papanicolaou. 1996. General asymptotic estimates for the coupon collector problem. *Journal of Computational and Applied Mathematics*, 67(2):277–289.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL2007)*, pages 858–867. ACL.

Adam L. Buchsbaum, Raffaele Giancarlo, and Jeffery R. Westbrook. 1998. Shrinking Language Models by Robust Approximation. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1998)*, pages 685–688.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL 1996)*, pages 310–318. ACL.

Ken Church, Ted Hart, and Jianfeng Gao. 2007. Compressing Trigram Language Models with Golomb Coding. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 199–207. ACL.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Jianfeng Gao and Min Zhang. 2002. Improving Language Model Size Reduction using Better Pruning Criteria. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 176–182. ACL.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2011. Producing Power-Law Distributions and Damping Word Frequencies with Two-Stage Language Models. *Journal of Machine Learning Research*, 12:2335–2382.

Joshua Goodman and Jianfeng Gao. 2000. Language Model Size Reduction by Pruning and Clustering. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 110–113. ISCA.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America (PNAS 2004)*, volume 101, pages 5228–5235.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. 2002. Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1–6. ACL.

Le Quan Ha, P. Hanna, D. W. Stewart, and F. J. Smith. 2006. Reduced n-gram models for English and Chinese corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (COLING-ACL 2006)*, pages 309–315. ACL.

Teemu Hirsimaki. 2007. On Compressing N-Gram Language Models. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pages 949–952.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28.

Hailong Li. 2005. On Generalized Euler Constants and an Integral Related to the Piltz Divisor Problem. *Šiauliai Mathematical Seminar*, 8:81–93.

Benoit B. Mandelbrot. 1965. Information Theory and Psycholinguistics: A Theory of Word Frequencies. In *Scientific Psychology: Principles and Approaches*. Basic Books.

OEIS. 2001. The on-line encyclopedia of integer sequences (a061017). `http://oeis.org/A061017/`.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*, pages 424–440. Lawrence Erlbaum Associates.

Andreas Stolcke. 1998. Entropy-based Pruning of Backoff Language Models. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274.

Frank Wood, Jan Gasthaus, Cédric Archambeau, Lancelot James, and Yee Whye Teh. 2011. The Sequence Memoizer. *Communications of the Association for Computing Machines*, 54(2):91–98.

George Kingsley Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin.