DINCON 2017
CONFERÊNCIA BRASILEIRA DE DINÂMICA, CONTROLE E APLICAÇÕES
30 de outubro a 01 de novembro de 2017 – São José do Rio Preto/SP

Estimação da entropia e as leis de Zipf e Heaps

Leonardo C. Araújo[1]
Universidade Federal de São João del-Rei
Hani C. Yehia[2]
Universidade Federal de Minas Gerais
Thaïs Cristófaro-Silva[3]
Universidade Federal de Minas Gerais

Resumo. Este trabalho utiliza a abordagem de Harris para calcular do valor esperado da polarização do estimador *plug-in* da entropia, analisando o caso de linguagens naturais, onde observa-se a presença da lei de Zipf e da lei de Heaps. Nos exemplos analisados, o valor esperado da polarização do estimador variou de 1,5 a 3,3% do valor da entropia estimada. Esta polarização não é desprezível, devendo ser considerada, uma vez que usualmente lidamos com dados empíricos que sofrem de má estimação, principalmente na cauda da distribuição, o que acarreta em um maior desvio na entropia estimada.

Palavras-chave. estimação de entropia, polarização do estimador, lei de Zipf, lei de Heaps.

1 Introdução

A incerteza associada a palavras em uma língua é um conceito teórico sobre a quantidade de informação produzida por uma fonte que utiliza um determinado arcabouço de símbolos. Este conceito é central na área da linguística quantitativa e computacional, com aplicações que vão desde análise de diversidade lexical, grau de diversificação do uso de palavras em uma língua, atribuição de autoria de textos, processamento de linguagem natural, dentre outras. O conceito de entropia foi introduzido por Shannon [1] como uma forma de medir informação ou incerteza. Para um alfabeto finito $\mathcal{X} = \{a_1, \ldots, a_N\}$, com uma distribuição subjacente $\{p_k\}_{k=1,\ldots,N}$, a entropia de Shannon é definida na forma $H = -\sum_k p_k \log p_k$, onde usualmente utiliza-se $\log = \log_2$ e, desta forma, a entropia é medida em *bits*.

Desde a definição formal da entropia, como uma forma de medir a incerteza ou a 'escolha' inerente à produção de símbolos por uma fonte, Shannon e outros pesquisado-

[1]leolca@ufsj.edu.br
[2]hani@ufmg.br
[3]thaiscristofarosilva@ufmg.br

res buscaram estimar precisamente a entropia do inglês escrito [2–4]. Outros trabalhos também foram desenvolvidos para estimar a entropia de outras línguas [5–8].

Entretanto, para conhecer a entropia de uma fonte, subentende-se que a sua distribuição é conhecida, o que, de forma geral, não ocorre. Para estimar a entropia através da simples contagem de ocorrências, devemos assumir que a ocorrência dos *tokens*[4] se dá de forma independente, ou seja, no caso de um texto estaremos supondo que não existe dependência entre palavras subsequentes, o que claramente não corresponde à realidade. Apesar dos contrapontos, é usual a utilização da estimativa de máxima verossimilhança para as probabilidades e a utilização do estimador *plug-in* para a entropia [9]. Dada uma amostra de comprimento M, $x_{1:M}$, a estimativa de máxima verossimilhança para a probabilidade de ocorrência do k-ésimo símbolo é dada por $\hat{p}_k = n(a_k|x_{1:M})M^{-1} = {}^{n_k}/_{M}$, onde $n_k{=}n(a_k|x_{1:M})$ denota o número de ocorrências do k-ésimo símbolo em uma dada amostra de comprimento M. O estimador *plug-in* da entropia é aquele em que utilizamos diretamente as estimativas das probabilidades dos símbolos, como se elas realmente constituíssem a real distribuição da fonte, ou seja, $\hat{H} = -\sum_k \hat{p}_k \log \hat{p}_k$ [9].

O problema de estimação é um problema antigo e difícil. A estimação da entropia lida inerentemente com ambos problemas de estimação: da distribuição e da entropia, em si. A medida de entropia de Shannon é uma grandeza em que a contribuição dos eventos raros é preponderante, sendo mais sensível a pertubações nas probabilidades da cauda da distribuição. A entropia pode ser vista como o valor esperado da informação associada aos eventos. A variação da informação associada a um evento é inversamente proporcional à sua probabilidade. Desta forma, pequenas pertubações nas probabilidades dos eventos menos prováveis podem acarretar grandes variações na medida final de entropia. São justamente os eventos raros aqueles que possuem estimação mais pobre, em alguns casos, eventos extremamente raros sequer serão observados em nossa amostra.

Desde o século XVII é notória a preocupação com a estimação das probabilidades. Laplace [10] introduziu a regra de sucessão, lidando com o problema da estimação da probabilidade de que o sol irá nascer amanhã, tendo em vistas todas as observações passadas de que ele, de fato, nasceu. Ele introduziu a primeira técnica de ajuste das probabilidades das distribuições, técnicas estas que hoje são conhecidas como métodos de suavização (*smoothing*). Um século depois, este problema também foi abordado por Alan Turing, sendo sua proposta de ajuste sistematizada e publicada por seu colega, Irving John Good [11].

O estimador *plug-in* para a entropia, é um estimador não-paramétrico, prático, intuitivo, e cuja convergência é a mais rápida, dentre os estimadores não-paramétricos conhecidos [12]. Antos e Kontoyiannis [13] mostraram que, para todas as distribuições com entropia finita, a variância da estimação $\hat{H}$ possui um limite superior que decai com o comprimento da amostra M, a uma taxa de $\mathcal{O}(\ln^2(M)M^{-1})$, sendo o mesmo limite válido para todas distribuições em alfabetos infinito contáveis. Para alfabetos com cardinalidade conhecida, $|\mathcal{X}|{=}N$, a variância de $\hat{H}$ decai a uma taxa $\mathcal{O}(M^{-1})$, que não pode ser melhorada pois o estimador *plug-in* é o estimador de máxima verossimilhança quando o tamanho do alfabeto é conhecido. Entretanto, o estimador é polarizado, sendo a entropia

[4] *Tokens* são instâncias de um fenômeno ou tipo que ocorrem no corpus observado. Um tipo é uma classe a qual os *tokens* do mesmo tipo pertencem.

comumente sub-estimada, visto que os eventos menos prováveis são, de forma geral, subestimados. Harris [14] propôs a seguinte forma para o valor esperado da polarização da entropia estimada:

$$\mathrm{E}[\hat{H} - H] = -\frac{N-1}{2M} + \frac{1}{12M^2}\left(1 - \sum_{k=1}^{N} p_k^{-1}\right) + \mathcal{O}(M^{-3}). \tag{1}$$

Em muitos problemas, o próprio valor de N é desconhecido e também precisa ser estimado. O termo de ordem $1/M^2$ na Equação 1 ainda depende das probabilidades p_k, as quais usualmente não podem ser estimadas de forma confiável. Para valores grandes de N, em relação à amostra, o primeiro termo da Equação 1 é considerável. Harris ainda formulou a variância do estimador através da seguinte expressão:

$$\mathrm{Var}(\hat{H}) = \frac{1}{M}\left(\sum_{k=1}^{N} p_k \log^2 p_k - H^2\right) + \frac{N-1}{2M^2} + \mathcal{O}(M^{-3}). \tag{2}$$

É importante observar que as expressões obtidas por Harris fornecem o valor esperado e variância da estimação da entropia em *nats*, devendo ser feita a conversão para *bits* quando necessário for.

No presente trabalho iremos analisar a polarização da estimação da entropia para o caso em que a distribuição subjacente segue a lei Zipf e, consequentemente, a lei de Heaps. Em seguida, analisaremos os resultados para alguns textos em inglês e verificaremos o quão significativos são os desvios na estimação.

1.1 Lei de Zipf e Heaps

A lei de Zipf é uma regularidade controversa observada em diversas áreas. George Kingsley Zipf [15] fez suas observações analisando a frequência de ocorrência de palavras e constatou que existe uma relação entre esta e o seu ranque, o produto de ambas é aproximadamente uma constante. Desta forma, existe uma lei de potência que rege a distribuição de Zipf, $p_k = f(k; N, s) = H_{N,s} k^{-s}$, onde $H_{N,s}^{-1} = \sum_{n=1}^{N} n^{-s}$ é a constante de normalização, conhecida como número harmônico generalizado, k é o ranque associado a cada símbolo e s o expoente característico da distribuição. Esta regularidade já foi constatada em diversas línguas naturais e também em outros fenômenos naturais.

Uma outra relação de escala frequentemente observada em linguagens naturais é a lei de Heaps (também conhecida como lei de Herdan) [16, 17]. Ela aborda a forma como o vocabulário de uma língua cresce com uma função sublinear do tamanho da amostra observada. Ao realizar um gráfico logarítmico em ambas coordenadas, observamos aproximadamente uma reta, sugerindo uma relação de potência entre o tamanho do vocabulário, N, e o comprimento da amostra M, ou seja, $N(M) = \alpha M^{\beta}$, sendo $0<\beta<1$ para conferir o crescimento sublinear. A coexistência de ambas as leis foi verificada inicialmente no Inglês, Russo e Espanhol por Gelbukh e Sidorov [18], sendo matematicamente verificada por Leijenhorst et al. [19], para o caso em que $s>1$, levando a $\beta=1/s$. De fato, o expoente

de Zipf encontrado em linguagens naturais é maior do que 1[5]. Desta forma, podemos assumir que a lei de Heaps também estará presente nestes sistemas.

Aplicando o logaritmo a ambos os lados da relação dada pela lei de Heaps, obtemos $\log N = \beta \log M + \log \alpha$, onde β representa o coeficiente angular (expoente da relação de Herdan) e $\log \alpha$ o coeficiente de intercepção. Como, para uma amostra de comprimento M=1, deveremos ter N=1 e assim deveremos ter α=1. O outro parâmetro pode ser determinado por $\beta = {}^{\log N}/_{\log M}$. O parâmetro β, entretanto, é dependente do tamanho da amostra. Desta forma, não podemos confiar neste parâmetro como sendo uma constante característica, independente do comprimento da amostra [22]. Ao realizar uma regressão linear para ajuste dos dados, frequentemente encontraremos valores de $\alpha \neq 1$ [17].

2 Zipf e Heaps na estimação da entropia

Iremos aqui utilizar as relações dadas nas lei de Zipf e Heaps para efetuar o cálculo da polarização do estimador dado na Equação 1. Utilizando o valor das probabilidades dadas por Zipf, teremos que a polarização do estimador, $\mathrm{E}H_{\mathrm{bias}}$, poderá ser escrita como

$$\mathrm{E}[\hat{H} - H] = -\frac{N-1}{2M} + \frac{1}{12M^2}\left(1 - H_{N,s}^{-1} H_{N,-s}^{-1}\right) + \mathcal{O}(M^{-3}), \tag{3}$$

e utilizando agora a relação de Heaps teremos

$$\mathrm{E}H_{\mathrm{bias}} = -\frac{N-1}{2\left({}^{N}/_{\alpha}\right)^{s}} + \frac{1}{12\left({}^{N}/_{\alpha}\right)^{2s}}\left(1 - H_{N,s}^{-1} H_{N,-s}^{-1}\right) + \mathcal{O}(M^{-3}), \tag{4}$$

sendo esta agora uma expressão apenas de s, N e α.

A Figura 1 apresenta o comportamento do desvio esperado relativo na estimação da entropia, em função do expoente de Zipf s e do tamanho do alfabeto N. Este comportamento apresentado através da razão entre o desvio esperado e a entropia, evidenciando que o desvio é mais significativo quão menor for o expoente de Zipf. Linguagens naturais apresentam $s \approx 1$, desta forma podemos observar que a polarização do estimador não deve ser menosprezada, sendo esta crescente com o tamanho do alfabeto.

3 Resultados

Iremos analisar os resultados utilizando alguns textos obtidos no Projeto Gutenberg. Os textos foram processados computacionalmente através de uma combinação de *scripts* para *Bash*, *Python* e *GNU Octave*.

Convencionamos como unidade básica as palavras em uma língua, ou seja, as palavras serão consideradas como os símbolos produzidos pela nossa fonte. O expoente característico s da lei de Zipf foi estimado para os textos, conforme descrito em [21]. A entropia da fonte foi estimada através de três métodos: 1) utilizando o estimador *plug-in*, onde

[5] Valores encontrados para o expoente de Zipf menor do que 1, usualmente decorrem de erro na estimação quando o comprimento da amostrada não é grande o suficiente. Este problema é amenizado quando consideramos o modelo generalizado proposto por Mandelbrot [20, 21].

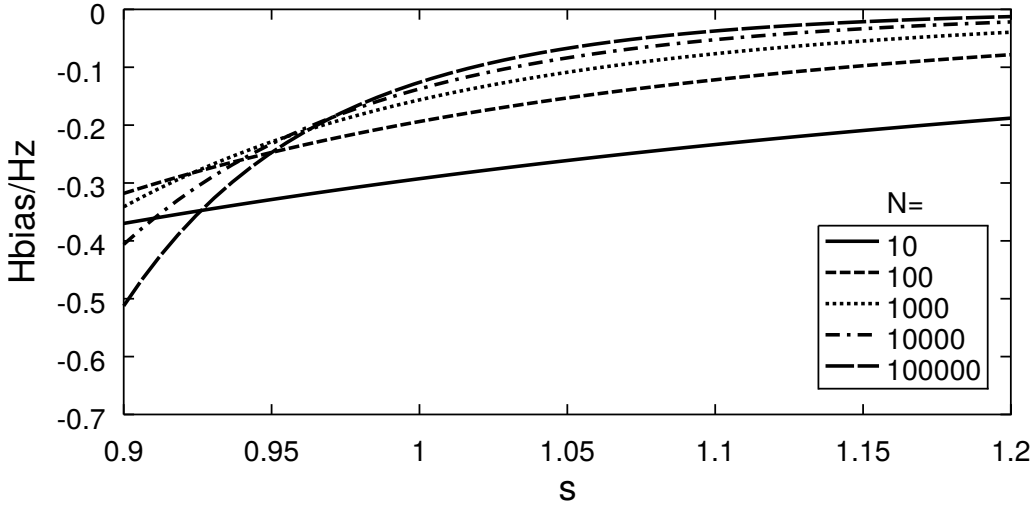

Figura 1: Relação entre a polarização da estimação da entropia e a entropia na presença da lei de Zipf e Heaps.

foram utilizadas estimativas de máxima verossimilhança das probabilidades [9]; 2) utilizando o estimador *plug-in*, com as probabilidades suavizadas através do método *Simple Good-Turing* (SGT) [23]; 3) utilizando a correção de Miller-Madow [14, 24]; 4) utilizando o método *Jackknife* [25]; 5) e a polarização dada na Equação 4. Os resultados são apresentados na Tabela 1.

Tabela 1: Entropia estimada para textos (bits). A tabela apresenta o tamanho do alphabeto N (número de tipos); o tamanho da amostra M (número de tokens); o valor do expoente de Zipf estimado $\hat{s}$; a entropia estimada utilizando o estimador de máxima verossimilhança (plug-in) H_{mle}; a entropia estimada com a correção de Miller-Madow; a entropia estimada pelo método Jackknife; e a polarização da estimação, dada pela Equação 4. Para os campos onde há a divisão entre 'obs' e 'sgt', utiliza-se, no primeiro, os números de observações na dada amostra e, no segundo, a suavização dada pelo método SGT.

fonte	N	M	$\hat{s}$		$\hat{H}_{\mathrm{mle}}$		H_{mm}		H_{jk}		$\mathrm{E}H_{\mathrm{bias}}$
			normal	sgt	normal	sgt	normal	sgt	normal	sgt	
Alice	2860	27807	1,189	1,356	8,737	8,456	8,788	8,511	8,849	8,366	-0,193
Hamlet	4827	32035	1,027	1,185	9,346	8,953	9,421	9,035	9,525	8,727	-0,222
Macbeth	3546	17751	0,962	1,117	9,467	9,045	9,567	9,156	9,707	8,751	-0,291
Shakespeare	26949	888113	1,461	1,632	10,010	9,867	10,025	9,882	10,042	9,831	-0,255
Ulysses	29845	264626	1,033	1,192	10,624	10,218	10,681	10,278	10,755	10,024	-0,159

4 Conclusões

A entropia de Shannon é uma função da distribuição da fonte, desta forma, conhecer a entropia implica em conhecer as características estatísticas subjacentes da fonte.

Ao analisar sinais e fenômenos da natureza, estas informações não nos são aprazíveis. Portanto, modelos e estimações devem ser utilizados, visando aproximar as características observáveis dos nossos dados. Por melhor que sejam, modelos e estimações estarão sempre sujeitos imperfeições. Neste trabalho buscamos revisitar a abordagem de Harris para averiguar a polarização do estimador *plug-in* para a entropia. Certos fenômenos da natureza apresentam características estatísticas que são bem descritas pela lei de Zipf. Em especial, as linguagens naturais são exemplos de fenômenos onde o comportamento descrito pela lei de Zipf é evidenciado. A lei de Heaps também foi utilizada, aproveitando a relação existente entre essas duas leis amplamente conhecidas na área da linguística quantitativa. Desta forma, foi realizada a análise dos modelos e estimações para alguns dados textuais obtidos livremente no Projeto Gutenberg. Aplicamos os resultados de Harris especificamente na condição em que a distribuições subjacente tem o comportamento dado pela lei de Zipf, a fim de verificar a polarização do estimador neste modelo. O estimador *plug-in* usualmente subestima a entropia, uma vez que a entropia é uma função na qual a cauda da distribuição exerce papel predominante e, em geral, as probabilidades associadas a eventos de menor probabilidade são mal estimadas. Para fontes com distribuição de Zipf, o expoente característico e, principalmente, o tamanho do vocabulário exercem influência na polarização do estimador, de forma que, quão maior forem estes parâmetros, menor será o erro cometido na estimação da entropia. A polarização do estimador é extremamente sensível ao expoente de Zipf, sendo desta forma preponderante utilizar um boa estimação de $\hat{s}$. Como, em linguagens naturais, o expoente característico é próximo de 1 e o tamanho do vocabulário, usualmente, menor que 10^5 (100 mil palavras)[6], teremos um erro na estimação da entropia menor que 2 bits. Um erro desta magnitude é significativo, uma vez que a entropia estimada para linguagens naturais usualmente encontra-se entre 4 e 10 bits por palavra [8].

Referências

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, 1948.

[2] ——, "Prediction and entropy of printed english," *Bell System Technical Journal*, 1951.

[3] P. F. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, and J. C. Lai, "An estimate of an upper bound for the entropy of english," *Computational Linguistics*, vol. 18, no. 1, pp. 31–40, 1992.

[4] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1319–1327, 2006.

[6] Utilizamos aqui uma definição simplificada de palavra, onde qualquer sequência de caracteres separada por espaços constitui uma palavra. Embora imperfeita, esta é uma simplificação para tornar o problema tangível. Todos os diferentes lexemas de um mesmo lema serão contabilizados como palavras distintas. Por outro lado, palavras compostas não serão contabilizadas, apenas as suas palavras componentes.

[5] F. H. Behr, V. Fossum, M. Mitzenmacher, and D. Xiao, "Estimating and comparing entropy across written natural languages using ppm compression," Harvard University, Tech. Rep., 2002.
[6] M. A. Montemurro and D. H. Zanette, "Universal entropy of word ordering across linguistic families," *PLoS ONE*, vol. 6, no. 5, p. e19875, 05 2011.
[7] ——, "Complexity and universality in the long-range order of words," *arXiv*, 2015.
[8] C. Bentz and D. Alikaniotis, "The word entropy of natural languages," *arXiv*, 2016.
[9] Z. Zhang and X. Zhang, "A normal law for the plug-in estimator of entropy," *IEEE Trans. Information Theory*, vol. 58, no. 5, pp. 2745–2747, 2012.
[10] P.-S. Laplace, *Essai philosophique sur les probabilités*. Paris: Paris Bachelier, 1840.
[11] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40(3-4), pp. 237–264, 1953.
[12] Z. Zhang, "Entropy estimation in turing's perspective," *Neural Computation*, vol. 24, no. 5, pp. 1368–1389, 2012.
[13] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures and Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.
[14] B. Harris, "The statistical estimation of entropy in the non-parametric case," University of Wisconsin-Madison, Tech. Rep. 1605, 1975.
[15] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner Pub. Co, 1949.
[16] G. Herdan, *Type-token Mathematics: A Textbook of Mathematical Linguistics*. Mouton en Company, 1960.
[17] H. S. Heaps, *Information retrieval. Computational and theoretical aspects*. New York: Academic Press, 1978.
[18] A. Gelbukh and G. Sidorov, "Zipf and heaps laws' coefficients depend on language," in *Computational Linguistics and Intelligent Text Processing: Second International Conference, CICLing 2001, Mexico-City, Mexico, February 18-24, 2001*. Springer Berlin Heidelberg, 2001, pp. 332–335.
[19] D. C. v. Leijenhorst, T. P. v. d. Weide, and F. Grootjen, "A formal derivation of Heaps' law," *Information Sciences*, vol. 170, no. 2-4, pp. 263–272, 2005.
[20] B. B. Mandelbrot, "Information theory and psycholinguistics," in *Scientific Psychology: Principles and Approaches*, B. B. Wolman and E. N. Nagel, Eds. Basic Books Publishing, 1965, pp. 550–562.
[21] L. C. Araujo, H. C. Yehia, and T. Cristófaro-Silva, "Entropy of a zipfian distributed lexicon," *Glottometrics*, vol. 26, pp. 38–49, 2013.
[22] R. H. Baayen, *Word Frequency Distributions*. Dordrecht: Kluwer, 2001.
[23] W. A. Gale and G. Sampson, "Good-turing frequency estimation without tears," *Journal of Quantitative Linguistics*, vol. 2, pp. 217–237, 1995.
[24] G. A. Miller, "Note on the bias of information estimates," pp. 95–100, 1955.
[25] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, vol. 9, no. 3, pp. 586–596, 1981.