现代汉语熵的计算及语言模型中稀疏事件的概率估计

黄萱菁,吴立德,郭以昆,刘秉伟
(复旦大学计算机科学系,上海 200433)

摘 要: 本文在大规模语料的基础上,利用语言模型中稀疏事件的概率估计方法对汉语的熵进行计算,并讨论了语料规模等因素对熵的影响. 在 4 年的人民日报的语料规模下,所求得的零阶熵、一阶熵、二阶熵分别为 9.62,6.18 和 4.89 比特.

关键词: 熵; 困惑度; 统计语言模型
中图分类号: TP391, G354 **文献标识码**: A **文章编号**: 0372-2112 (2000) 08-0110-03

Computation of the Entropy of Modern Chinese and the Probability Estimation of Sparse Event in Statistical Language Model

HUANG Xuan-jing, WU Li-de, GUO Yi-kun, LIU Bing-wei
(*Dept. Of Computer Science, Fudan University, Shanghai* 200433, *China*)

Abstract: Different estimation methods of the probabilities of sparse events for the computation of the entropy in large-scale modern Chinese text are applied in this paper. Experiments based on the corpus of four years' People's Daily show that the 0-order, 1-order and 2-order entropy are 9.62, 6.18 and 4.89 bits respectively. In addition, the influence of such factors as the scale of corpus is also discussed.

Key words: entropy; perplexity; statistical language model

1 引言

近年来,统计语言模型逐渐在自然语言处理和语音处理中得到了广泛的应用. 为了比较不同的语言模型的差异性,必须考察各种模型的不确定性。模型的不确定性越大,正确估计语言现象的可能性就越小. 语言模型的不确定性可用"熵"来加以定量的描述. 在英语中,若以词汇作为统计的基本单位,则 1-gram、2-gram 和 3-gram 的熵分别为 11.47、6.06 和 2.01 比特[1]. 对于汉语的熵,不少学者也利用不同的方法得到了各自的数值:石贵青和徐秉铮等人计算出汉字的 0 阶熵为 9.42 比特,通过猜字实验估算出汉字的熵不大于 4.1 比特[2,3]. 吴军、王作英根据统计方法计算出的熵为 5.17 比特[4]. 冯志伟根据英汉双语语料库的对比研究,测得汉字的极限熵平均值为 4.0462 比特[5].

由于计算机可读文本的大量出现,以及计算能力的不断提高,使得在更大的语料规模上,更为精确地用统计方法计算汉语的熵成为可能. 本文将在大规模文本的基础上,运用语言模型中稀疏事件的概率估计方法来计算汉语的熵,并讨论文本规模和可计算性等各种因素对熵的影响.

2 基本概念

本节将介绍语言模型中的几个概念,包括熵、交叉熵和困惑度.

语言的熵:语言的熵反映语言中每个字符的平均信息量. 对于语言 L,用 $x_1^n = x_1, \cdots x_n$ 表示 L 中长为 n 的语句,$p(x_1^n)$ 是 L 中 x_1^n 的概率. 根据信息论原理,若语言 L 是各态遍历的、平稳的随机过程,则有:$H(L) = -\lim_{n\to\infty}(1/n)\log p(x_1^n)$.

在通常情况下,分布 $p(x)$ 是未知的,只能通过某种估计方法建立模型 $q(x)$ 来近似 $p(x)$. 在统计自然语言处理中最常用的模型是 N 元(Ngram)模型,即 $N-1$ 阶的 Markov 模型.

交叉熵:为了评估模型 $q(x)$ 和真实分布 $p(x)$ 的的相似程度,引进了交叉熵的概念. 若语言 L 是各态遍历的、平稳的随机过程,而模型 q 是 N 元模型,用 $q(x_1^n)$ 表示模型 q 对 x_1^n 的概率估计,用 LL 表示训练文本的容量,则有:

$$H(L,q) \approx -\frac{1}{LL-N+1}\sum_{l=N}^{l=LL}\log q(x_l \mid x_{l-N+1}^{l-1}).$$

困惑度:语言 L 的模型 q 的困惑度 PP_q 为:$PP_q = 2^{H(L,q)} \approx 2^{-(1/LL)\log q(x_1^{LL})}$. 在利用了语言模型的情况下,给定一段历史,那么当前字符平均只可能有 PP_q 种选择. 这样,语言模型设计的任务就是寻找困惑度最小的模型,因为它最有效,也最接近真实的语言.

收稿日期:1999-01-29;修回日期:2000-02-29
基金项目:国家自然科学基金(No. 69873011);863 计划(No. 863-306-ZD02-02-4),上海市教委重点学科基金

3　稀疏事件的概率估计

在计算熵的时候,会不可避免地遇到数据稀疏的问题. 为此引入了许多稀疏事件的概率估计方法[6]. 其中,Katz 的回推方法是目前最好的,它假设未现事件 (h, w) 的概率与单字概率 $q(w)$ 和历史概率 $q(h)$ 的乘积成正比[7].

回推方法的计算量太大了. 由于我们必须估计每个未现事件的概率,所以整个估计算法的数量级为 $O(K)$,K 为全部可能的 Ngram 的个数. 这里提出了一个较为简化的估计方法以估计未现事件并进而计算熵,计算公式如下:

$$p(x_1, x_2, \cdots, x_N) = \begin{cases} (r - d_r)/NN, & r > 0 \\ \lambda_0 p(x_1)\, p(x_2) \cdots p(x_N), & r = 0 \end{cases};$$

这里 r 是 $x_1 x_2 \cdots x_N$ 的计数,λ_0 是归一化常数.

利用上式进行熵的计算是非常便利的. 用 seen 和 unseen 分别表示观察到的和未观察到的事件,于是有:

$$\begin{aligned} E(x_1 x_2 \cdots x_N) &= -\sum p(x_1 x_2 \cdots x_N) \log(p(x_1 x_2 \cdots x_N)) \\ &= -\sum_{seen} p(x_1 x_2 \cdots x_N) \log(p(x_1 x_2 \cdots x_N)) \\ &\quad -\sum_{unseen} p(x_1 x_2 \cdots x_N) \log(p(x_1 x_2 \cdots x_N)) \end{aligned}$$

式中的前一部分可直接计算. 对后一部分,有:

$$\sum_{unseen} p(x_1 x_2 \cdots x_N) \log(p(x_1 x_2 \cdots x_N)) = \lambda_0 \sum_{unseen} p(x_1) \cdots p(x_N) \cdot (\log(\lambda_0 p(x_1) \cdots \log p(x_N))) = \lambda_0 \log \lambda_0 - N \times \lambda_0 \times E(x_1) - \lambda_0 \sum_{seen} p(x_1) \cdots p(x_N) (\log(\lambda_0 p(x_1) \cdots \log p(x_N))) = \lambda_0 \log \lambda_0 - N \times \lambda_0 \times E(x_1) - \lambda_0 \sum_{seen} p(x_1) \cdots p(x_N) (\log(\lambda_0 p(x_1) \cdots \log p(x_N)))$$

上式的后一部分可直接计算;而 $\lambda_0 = \lambda / (1 - \sum_{seen} p(x_1)\, p(x_2) \cdots p(x_N))$,其中 λ 为折扣概率之和,由折扣函数决定.

因此上述算法的复杂度和可观察到的事件个数在同一数量级. 由于全部可能的 Ngram 中,只有很少的一部分可被观察到,因此算法的运行时间就大大缩短了.

4　实验结果及分析

4.1　语料库简介

实验所用的语料为人民日报 1993～1996 年的全文文本,共约 184 兆,7000 多万字(不包括标点符号在内). 这样规模的语料仍然只包含了所有可能的 Ngram 中极少的一个部分. 例如,可观察到的 4-gram 就只占全部的 0.00000092 %.

图 1 反映了每年新增加的 Ngram 在这一年全部的 Ngram 中所占的比例. 从图 1 中可以看出,Ngram 的增量呈明显减少的趋势. 这说明当语料进一步增长的情况下,Ngram 的总数有可能维持在一个较为稳定的水平.

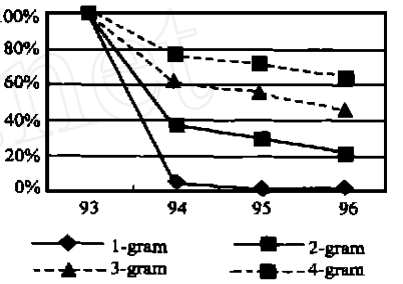

图 1　新增 Ngram 的百分比

4.2　熵的计算

在下面的实验中,对比了 7 种不同的概率估计方法,包括:最大似然估计、加一估计、减一估计、绝对折扣模型、线性折扣模型、留一估计和回推估计. 以 93 和 94 年的文本为训练集,以 95 和 96 年的文本为测试集. 所运用的评价标准是各种模型的困惑度,包括在训练集和测试集上的困惑度,分别用 PP_{train} 和 PP_{test} 表示.

各种方法所得到的熵和困惑度的值见表 1. 由于 Ngram 模型对应着 $N-1$ 阶的 Markov 模型,所以将相应的熵称为 $(N-1)$ 阶熵. 对 1-gram、2-gram 和 3-gram,给出了熵(单位:比特,以下同)和困惑度的值,而对于 4-gram,则只提供了熵,其原因将在下文介绍.

表 1　各种方法的熵和困惑度

方法	1-gram			2-gram			3-gram			4-gram
	0 阶熵	PP_{train}	PP_{test}	1 阶熵	PP_{train}	PP_{test}	2 阶熵	PP_{train}	PP_{test}	3 阶熵
最大似然估计	9.62	785.8	787.4	6.09	70.2	65.0	3.90	14.0	8.7	1.99
加一估计	9.62	784.2	788.0	14.68	264.5	295.1	15.03	91532.7	61572.6	13.11
减一估计	9.62	786.3	788.0	6.43	77.3	92.7	7.61	76.4	64.5	11.76
绝对折扣模型	9.62	785.8	788.0	6.31	72.3	88.8	6.68	17.3	33.1	10.18
线性折扣模型	9.62	785.8	788.0	6.37	71.4	88.6	7.06	16.1	31.6	10.66
留一估计($R0=6$)	9.62	786.9	788.0	6.22	71.5	88.8	6.31	16.7	30.4	10.04
回推估计($R0=6$)	9.62	786.9	788.0	6.14	71.5	84.6	5.02	16.7	30.2	6.59

从表 1 我们可以发现:

(1) 通过最大似然估计求得的熵最小. 这是因为训练集和测试集的 Ngram 交集很小,从而发生了明显的过分拟合现象;

(2) 加一和减一估计两种经验方法结果较差;

(3) 几种折扣方法,包括绝对折扣模型、线性折扣模型和留一估计其性能没有特别显著的差异,以留一估计略好,绝对折扣次之;

(4) 除了回推估计方法以外,其它方法所求得的 3-gram 的熵和困惑度均大于 2-gram. 这说明回推估计方法对未现事件的估计较为准确.

现在让讨论参数估计的可信度的问题. 对于 1-gram,训练集和测试集的困惑度非常接近,说明从现有的训练语料可较为准确地估计语言模型. 而对于 2-gram,两种困惑度之间已有一定的差距;对于 3-gram,两者的差距进一步扩大. 对于 4-

gram,由于各种方法计算出的熵都大于 3-gram 的熵,再计算它们的困惑度已经没有什么意义了.

接下来再讨论语料规模对熵的计算的影响. 表 2 给出了在不同的语料规模下所求得的 0～3 阶熵的数值. 从表中可以发现,语料规模对于 0 阶熵和 1 阶熵没有显著的影响;而随着语料规模的扩大,各种方法所得到的 2 阶熵和 3 阶熵,一致地呈现单调下降的趋势. 因此可以认为,在当前规模的语料下所得到的 0 阶熵和 1 阶熵的值已经是较为准确的,而随着语料规模的增长,更高阶的熵的值将逐渐下降并趋于稳定.

表 2 各种语料规模下的熵

Ngram	年份	减一估计	绝对折扣	线性折扣	留一估计 ($R_0=6$)	回推 ($R_0=6$)
0 阶熵	93	9.62	9.61	9.62	9.62	9.62
	93～94	9.62	9.62	9.62	9.62	9.62
	93～95	9.62	9.62	9.62	9.62	9.62
	93～96	9.62	9.62	9.62	9.62	9.62
1 阶熵	93	6.53	6.38	6.45	6.24	6.11
	93～94	6.43	6.31	6.37	6.22	6.14
	93～95	6.39	6.31	6.35	6.25	6.18
	93～96	6.36	6.29	6.33	6.23	6.18
2 阶熵	93	8.30	7.20	7.60	6.81	5.20
	93～94	7.61	6.68	7.06	6.31	5.02
	93～95	7.19	6.39	6.74	6.03	4.95
	93～96	6.89	6.16	6.50	5.84	4.89
3 阶熵	93	12.54	10.94	11.41	10.91	7.09
	93～94	11.76	10.18	10.66	10.04	6.59
	93～95	11.25	9.70	10.19	9.50	6.28
	93～96	10.79	9.28	9.77	9.03	6.02

5 结论

综上所述,本文在大规模语料的基础上,采用留一估计和回推估计相结合的方法对汉语的熵进行计算,在 4 年的人民日报的语料规模下,所求得的零阶熵、一阶熵、二阶熵分别约为 9.62,6.18 和 4.89 比特. 虽然这里所求得的熵仍然只是汉语熵的一个上界,而不是上确界,但和采用猜字的心理学方法得到的值已经是相当接近了.

参考文献:

[1] Bell. T. C. ,Text Compression [M]. Prentice Hall ,1990.

[2] 石贵青,徐秉铮. 汉字字频分布、最佳编码与输入问题 [J]. 电子学报,1984,12(4):94 - 96.

[3] 徐秉铮,吴立忠. 中文文本压缩的 LZW 算法 [J]. 华南理工大学学报(自然科学版),1989,17(3).

[4] 吴军,王作英. 汉语信息熵和语言模型的复杂度 [J]. 电子学报,1996,24(10):69 - 71.

[5] 冯志伟. 关于汉字的熵和极限熵致编辑部的一封信 [J]. 中文信息学报,1998,12(1):63 - 64.

[6] 吴立德. 大规模中文文本处理 [M]. 上海:复旦大学出版社,1997.

[7] Katz ,S. M. Estimation of probabilities from sparse data for the language model component of a speech recognizer [J]. IEEE Trans. Acoust , Speech ,Signal Processing ,ASSP-35 (1987):400 - 401.

作者简介:

黄萱菁 1972 年生,1998 年毕业于复旦大学计算机系,获博士学位,现留校任教. 目前主要从事自然语言处理和多媒体检索等方面的研究.

吴立德 1937 年生,1958 年毕业于复旦大学数学系. 教授,博士生导师. 目前主要从事自然语言处理、图像处理和计算机视觉等方面的研究.