汉语统计语言模型训练样本容量的定量化度量

张仰森
(北京信息科技大学智能信息处理研究所 北京 100192)

摘 要 统计语言模型的参数训练是语言建模的关键,选择多大的训练样本就能够达到建模的参数估计误差要求,是语言建模理论关心的问题之一。应用数理统计理论对汉语统计语言模型的训练语料样本容量进行了定量化描述,给出了汉语n-gram模型训练样本容量下界的估算方法及量化估算公式,可根据模型参数估计的误差要求计算出模型训练所需的样本容量。

关键词 汉语统计语言模型,训练语料样本,样本容量,相对误差
中图法分类号 TP391 **文献标识码** A

Quantitative Measurement of Training Sample Capacity for Chinese Statistical Language Model

ZHANG Yang-sen
(Institute of Intelligent Information Management, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract The training of statistical language model parameter is the key of language modeling. Chooseing how many training samples to meet the demand of the model parameter estimation error is one of concern problems of language modeling theory. We applied mathematical statistics theory to give the estimating method for training samples lower bound capability for chinese model, the quantification estimation formula was suggested. By using this formula, the corpus sample capability needed to train model parameters can be calculated according to the demand of parameter estimation error.

Keywords Chinese statistical language model, Training corpus sample, Sample capacity, Relative error

1 汉语语言模型训练样本的选择

构建统计语言模型时非常重要的一步是对语言模型的参数进行训练,以便语言模型能够很好地符合真实语言的内在规律。但不可能应用真实语言中的全体语句对语言模型进行训练,只能抽取一部分具有代表性的语料,这些有代表性的语料称作样本,选取样本的过程称作抽样。

训练样本的选取是一个十分重要的过程,因为样本应该能够代表语言文本的全体,若选取得好,所训练出的语言模型就能够很好地表述语言的内在规律;反之,所得到的语言模型在应用中就会出现很大的误差,基于这样的语言模型所建立的应用程序或软件将无任何应用价值。

语料样本的选取应该考虑如下几个因素:

(1)语料样本的覆盖面。人们希望所建立的语言模型能够适应于语言所能描述的各个领域,因此,样本的选取应具有广泛的覆盖性,所谓的平衡语料库就是人们试图得到具有广泛覆盖性样本的一种努力。但就目前自然语言处理的技术而言,建立一个广谱的或万能的语言模型几乎是不可能的。面向应用或面向领域的语言模型的训练样本选择相对来讲要容易些。

(2)样本语料的规模。样本语料的规模与对模型参数训练的误差要求密切相关,也和建立语言模型时所使用的语言单位在语言应用中的使用频率密切相关。若要求模型参数的相对误差越小,则需要的训练语料样本的规模越大;而在一定的相对误差要求下,若被统计的语言单位所出现的频率越小,则需要的训练语料样本的规模也越大。

在以上两个因素中,有关语料样本的覆盖面问题,由于目前所建立的语言模型是面向领域或应用的,因而易于解决;样本语料规模的大小关系到能否快速有效地训练模型,是语言模型构建过程中迫切需要解决的问题。

2 统计语言模型阶数与训练样本规模的关系

统计语言模型中,n-gram模型是最常见的一种,其构造主要由两个因素决定:模型的阶数 n 和构造模型的基本单元。显然模型的阶数越高,对语言的描述越贴切,性能也就越好,但实现起来就越困难,因为模型阶数越高,数据的稀疏性就越严重,所需要的训练语料样本的规模也就越大。比如说,建造Trigram模型比建造Bigram模型就需要更多的训练语料,建造Bigram模型就比建造unigram需要更多的训练语料。建模单元的大小也对训练语料的规模有很大的影响,例如,建立汉语语言模型时,以字为单位的语言模型的训练语料规模就小于以词为单位建立的语言模型所需的语料规模。那么,建

到稿日期:2008-11-14 返修日期:2009-02-05 本文受国家自然科学基金(60873013),北京市自然科学基金B类重点项目(KZ200811232019),中国科学院自动化研究所模式识别国家重点实验室开放基金,北京市属市管高校人才强教计划项目资助。
张仰森(1962—),男,博士,教授,CCF会员,研究方向为中文信息处理、人工智能。

造 n-gram 汉语统计语言模型到底需要多大规模的训练语料呢？这是计算语言学界广大研究人员关心和研究的问题。文献[1]通过研究现代汉语中 n-gram 个数-频次关系，给出了一个判定 n-gram 训练充分性的定量判定标准：n-gram 模型个数-频次关系曲线与其相应的线性拟合曲线的标准差是衡量该统计语言模型训练充分性的定量标准。标准差越小表明该模型的训练量越充分；反之，表明该模型的训练量欠缺。这是一个非常有意义的结果，对汉语语言模型的构建与训练具有重要的指导意义。利用这种方法给出的是训练语料规模的上限估计，计算工作量也比较大。本文希望运用数理统计理论，通过数学推理方法给出 n-gram 模型训练语料规模的必要性量化度量方法，即推理出训练语料规模的下限。

前面已经指出，对于统计 n-gram 模型而言，其阶数 n 无疑是影响它的性能的最为重要的参数之一，高阶 n-gram 模型之所以难以构造，主要在于其参数估计时所需要的训练语料规模要大于低阶 n-gram 所需要的训练语料规模。显然，建造 unigram 模型所需的训练语料规模应该最小，也就是说，训练 unigram 模型的语料规模应是 n-gram 模型训练所需语料规模的下限。

Unigram 模型又称上下文无关(context-free)模型，它在估算当前词出现的概率时，并不考虑该词所在上下文环境对它的出现概率的影响。这是一种最简单也最易于实现的统计语言模型，因为它所需要的训练语料最少。它可以看作是建立高阶、实用的统计语言模型的基础。最常用的线性插值法和回退法建造统计语言模型的技术，在出现数据稀疏时所进行的模型平滑或回退都要用到 Unigram 模型。Unigram 模型实现时有两种方法：一种是将语言中的所有语言单位(词或字，以后假设为词)视为具有相同概率分布，这种情况下，一个词在文本中的出现概率由下式给出：

$$p(w_i|h)=p(w_i)=\frac{1}{|V|} \tag{1}$$

其中，h 表示 w_i 的上下文，$|V|$ 表示语言词典中词的数量。这样的模型显然是最简的，但却是最不实用的，因为假设所有词具有相同的概率分布不符合实际。因此通常人们所说的 Unigram 模型是根据所有词在语言中出现的概率不等这一事实，使用最大似然估计法(MLE)，通过训练文本中词的出现频次来估算词的出现概率。这样得到的语言模型能够更精细地反映词的统计特征，由于不考虑其上下文影响，一个词在文本中的出现概率可表示如下：

$$p(w_i|h)=p(w_i) \tag{2}$$

对于概率的估计通常采用 MLE 方法由下式近似地给出：

$$p(w_i)\approx\frac{count(w_i)}{N} \tag{3}$$

其中，$count(w_i)$ 表示词 w_i 在训练文本中出现的频次，N 为训练文本的总词数。

式(3)是词 w_i 的统计频率，它是对词 w_i 在语言中真正使用频率进行模拟的模型。可以看出，词在语料中出现频次 $count(w_i)$ 的统计精确性以及语料规模 N 对模型参数的估计是极其重要的。词是组成自然语言最基本的语言单位之一，且数量巨大，不同的词使用的频率相差很大。在汉语中，大家普遍能够认同(因为中文词的定义至今没有一个公认的标准)的词的数量也有 5,6 万之多，而这么多的词在通常的文本中的使用机会也是很不相同的，受人们的文化习惯、社会环境、专业领域甚至包括语法在内各种因素影响，有些词被使用的机会就非常多，而有些词的使用机会就比较少，甚至几乎不用。例如，根据对 1995 年人民日报的统计，汉语中虚词“的”、实词“中国”的频率就比较高，而有些词如“芒果”、“板蓝根”、“养鱼业”只出现了一次，有些专业性较强的词根本就不会出现。因此，建立语言模型时既要注意语料的选择覆盖面，同时还要扩大规模，本文受随机抽样中样本容量估计确定方法[2-5]的启发，试图根据汉语词的使用频率的粗略估计以及统计误差的要求，对汉语 n 元模型的训练样本规模做粗略估计。因此，为了后面叙述的方便，先给出以下定义。

定义 1 在自然语言中，词被使用机会的大小称作词的使用频率。

显然，一个自然语言中，要想确切地知道其每个词的固有使用频率几乎是不可能的，因为自然语言本身是一个随机变化的动态过程。因此，词的使用频率一般通过统计方法近似求得。为此，给出下列定义。

定义 2 利用统计方法求得某词在语料中的出现次数称为该词的频次；词的频次与统计文本总词数的比称为词的统计频率。

当统计语料规模达到一定量后，统计频率就基本趋于稳定，可以作为词在某种语言中的使用频率的近似度量。然而，在对语言模型进行训练时，由于词的数量巨大，要想得到词表中所有词的统计频率，就需要对世界上所有该语言的相关文本进行统计，但这实际上是不可能实现的。现只能抽取一部分文本资料进行统计，在满足一定误差要求的情况下得到词的统计频率。

定义 3 为获取词的统计频率而选取的含有 N_w 个词的文本资料作为统计样本语料，称 N_w 为训练语料规模或训练样本容量。

3 汉语语言模型的训练语料规模估计

在式(3)所表示的 Unigram 模型中，所要求取的就是 w_i 的使用频率，式(3)的右边则是该词的统计频率。下面就从式(3)出发，求取 n-gram 训练样本语料的规模下界。

设 w_i 是某一个词，其在语言中的使用频率为 p，由于语言的随机性，p 的值无法准确知道而有待于估计。假设 T 是以合理的方式选择语言模型训练语料样本，其中含有的词的个数为 N_w，若以 X 表示词 w_i 在 T 中的出现次数，则根据随机变量的概念，可以将 X 看作为一个随机变量，且 X 服从于参数为 N_w 和 p 的二项分布：即 $X\sim B(N_w,p)$，由定义 2 可以得到词 w_i 的统计频率为 $\frac{X}{N_w}$，它也是一个随机变量。根据伯努利大数定理[6]，则存在任意正数 $\varepsilon>0$，有：

$$\lim_{N_w\to\infty}P\left(\left|\frac{X}{N_w}-p\right|<\varepsilon\right)=1 \tag{4}$$

或

$$\lim_{N_w\to\infty}P\left(\left|\frac{X}{N_w}-p\right|\geqslant\varepsilon\right)=0 \tag{5}$$

即当训练语料的抽样规模 N_w 趋于无穷大时，$\frac{X}{N_w}$ 依概率收敛于词 w_i 的使用频率 p。但在实际中，训练语料的规模 N_w 不可能无穷大，确定训练语料规模 N_w 的目标是为了语言模型参数训练的简单实用，不要求式(4)中的 $|\frac{X}{N_w}-p|<\varepsilon$ 永远成立，只要求有很大的概率成立，以便使 N_w 的取值尽量的

小。为此，设 $0.5<\delta<1$ 表示很大的概率值，称为可信度，目标是使式(6)成立：

$$P\left(\left|\frac{X}{N_w}-p\right|<\varepsilon\right)\geqslant\delta \tag{6}$$

即，在给定允许误差 ε 和可信度 δ 的条件下，求出满足式(6)的最小 N_w 值。求解过程如下。

由于随机变量 X 服从于参数为 N_w 和 p 的二项分布：即 $X\sim B(N_w,p)$，因此，X 的数学期望 $E(X)$ 和方差 $D(X)$ 可表示为：

$$E(X)=N_wp \tag{7}$$

$$D(X)=N_wp(1-p) \tag{8}$$

词 w_i 的统计频率 $\frac{X}{N_w}$ 也是随机变量，根据数学期望与方差的性质[6]可知：

$$E\left(\frac{X}{N_w}\right)=\frac{1}{N_w}E(X)=p \tag{9}$$

$$D\left(\frac{X}{N_w}\right)=\frac{1}{N_w^2}D(X)=\frac{1}{N_w}p(1-p) \tag{10}$$

式(9)表明，应用 $\frac{X}{N_w}$ 来估计词 w_i 的使用频率是合理；式(10)表明训练语料的规模 N_w 越大，$\frac{X}{N_w}$ 的方差就越小，$\frac{X}{N_w}$ 逼近 p 时在其周围的摆动幅度就越小，$\frac{X}{N_w}$ 对 p 的近似度就越高，语言模型的描述能力越强。由中心极限定理[6]，$\frac{X}{N_w}$ 的标准化变量：

$$Y=\frac{X/N_w-p}{\sqrt{p(1-p)/N_w}} \tag{11}$$

的分布函数可以由标准正态分布函数 $\Phi(u)$ 来近似，即，当 N_w 充分大时，有

$$\frac{X/N_w-p}{\sqrt{p(1-p)/N_w}}\sim N(0,1) \tag{12}$$

其中，$N(0,1)$ 为标准正态随机变量的记号。对标准正态分布函数 $\Phi(u)$ 来说，可以利用标准正态分布表查出当 $u>0$ 时，随机变量落在区间 $(-\infty,u)$ 中的值为 $\Phi(u)$。对于某个负数 a，由标准正态随机变量的性质可知：$\Phi(a)=1-\Phi(|a|)=1-\Phi(-a)$。因此标准正态随机变量落在区间 $(-u,u)$ 概率为：

$$\Phi(u)-\Phi(-u)=2\Phi(u)-1 \tag{13}$$

将式(12)与式(6)做比较，对式(6)中的不等式 $\left|\frac{X}{N_w}-p\right|<\varepsilon$ 两边同除以 $\sqrt{p(1-p)/N_w}$，则式(6)变为：

$$P\left(\frac{|X/N_w-p|}{\sqrt{p(1-p)/N_w}}<\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}\right)\geqslant\delta \tag{14}$$

即：

$$P\left(-\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}<\frac{X/N_w-p}{\sqrt{p(1-p)/N_w}}<\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}\right)\geqslant\delta \tag{15}$$

由式(12)可知，式(15)左边的概率为标准化正态随机变量 Y 落在区间 $\left(-\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}},\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}\right)$ 的概率。将 $\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}$ 代入式(13)，则式(15)变为：

$$2\Phi\left(\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}\right)-1\geqslant\delta$$

即：$\Phi\left(\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}\right)\geqslant\frac{1+\delta}{2}$ (16)

当 $\frac{1+\delta}{2}$ 已知时，可通过标准正态分布表查出 $\Phi^{-1}\left(\frac{1+\delta}{2}\right)$，这时式(16)可变为：

$$\frac{\varepsilon\cdot\sqrt{N_w}}{\sqrt{p(1-p)}}\geqslant\Phi^{-1}\left(\frac{1+\delta}{2}\right) \tag{17}$$

对不等式(17)进行求解，即得满足式(6)误差要求的训练语料规模 N_w：

$$N_w\geqslant\frac{p(1-p)}{\varepsilon^2}\left[\Phi^{-1}\left(\frac{1+\delta}{2}\right)\right]^2 \tag{18}$$

此式即为所求得的 unigram 模型训练语料规模的下界表达式。它的估计受 3 个参数的影响，ε 为应用 MLE 方法从训练样本语料中估计词 w_i 使用频率 p 的误差要求，δ 是人们对满足误差要求的统计频率的可信程度，p 为建模语言单位在语言中的固有使用频率，它的值直接影响语料训练样本规模的估计。下面就讨论训练样本规模与词的使用频率和统计误差要求之间的关系。

4 训练样本规模与词的使用频率及统计相对误差之间的关系

在第 1 节已经指出，样本语料的规模与对模型参数训练的误差要求密切相关，也和样本中的统计语言单位的使用频率密切相关。如果要求的相对误差越小，所需要的样本规模就越大，因为若样本规模较小，样本中语言单位出现的频率就不够稳定，其与该语言单位在实际应用中的使用频率的误差就会很大。对于语言单位的频率而言，在给定误差要求的情况下，被统计语言单位的频率越低，所需要的样本规模就越大，因为若样本规模较小时，低频的语言单位出现的频率非常小甚至为 0，这就会导致估计误差达不到要求。

在自然语言中，高频词的出现频率与低频词的出现频率相差太大，对汉语语言来说，更是如此。显然，在应用统计频率近似使用频率时，对所有词使用式(6)显然是不太合理的，因为它将相同的绝对估计误差 ε 应用于所有的词。解决这一问题的办法是，对于出现频率低的词，将 ε 的值取得更小些，以保证估计的准确性。基于这一思想，可令 $\varepsilon=\alpha p$，把 ε 的取值与使用频率 p 联系起来，对使用频率低的词，误差要求就小。这里，α 为一介于 0 与 1 之间的小数，α 的值越小，统计频率对使用频率的估计就越准确。式(6)因此变成：

$$P\left(\left|\frac{X}{N_w}-p\right|<\alpha p\right)\geqslant\delta \tag{19}$$

这样做了之后，就可以将不同的估计误差要求应用于频率高低不同的词，特别是对低频词，这样的估计可能会更为理想。若取 $\alpha=1/3$，对于某类低频词，设其估计频率为 p_0，则 $\varepsilon=p_0/3$，将其代入式(18)，得训练语料规模：

$$N_w\geqslant\frac{9(1-p_0)}{p_0}\left[\Phi^{-1}\left(\frac{1+\delta}{2}\right)\right]^2 \tag{20}$$

例如，假设要建立基于字的 Unigram 汉语语言模型，建模的语言单位为字，这时可根据字的统计频率值，在给定可信度 δ 的情况下，获得该语言模型训练语料的最小规模要求。根据已有的字频统计资料[7]，假设取一类低频字的频率为

(下转第 249 页)

maintenance environment[C]//Proceedings of the ICDM Workshop on Integrating Data Mining and KnowledgeManagement. California,2001

[8] Missikoff M, Navigli R, Velardi P. Integrated approach to web ontology learning and engineering[J]. IEEE Computer,2002,35(11):60-63

[9] Budanitsky A, Graeme H. Evaluating WordNet-based measures of semantic distance[J]. Computational Linguistics, 2006, 32(1):13-47

[10] Cimiano P, Hotho A, Staab S. Learning concept hierarchies from text corpora using formal concept analysis[J]. Journal of Artificial Intelligence Research,2005(24):305-339

[11] Bisson G, Nedellec C, Canamero D. Designing clustering methods for ontology building: The Mo'K workbench[C]// Proceedings of the ECAI 2000 Workshop on Ontology Learning(OL'2000). 2000

[12] Maedche A, Staab S. Ontology learning[C]// Proceedings of 14th European Conference on Artificial Intelligence. 2000

[13] Faure D, Nedellec C. A corpus-based conceptual clustering method for verb frames and ontology acquisition[C]//Proc. LREC-98 Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, European Language Resources Distribution Agency. Paris,1998

[14] 孙亮,任小康. 基于本体的图像语义检索模型[J]. 重庆工学院学报:自然科学版,2009,23(1):127-131

(上接第 224 页)

2.6×10^{-6},即令 $p_0=2.6\times10^{-6}$,若这时令可信度 δ 分别取 0.93,0.95,0.97,0.99,则按式(20)计算得到表 1 所列出的结果。

表 1 基于字的 Unigram 模型训练样本规模

δ	0.93	0.95	0.97	0.99
$\Phi^{-1}(\frac{1+\delta}{2})$	1.81	1.96	2.17	2.58
N_w	1.13×10^7	1.33×10^7	1.63×10^7	2.30×10^7

表 1 表明,当训练语料规模达到 1130 万时,以式(3)所表示的 Unigram 模型,对所描述语言的估计有 93%的把握能够达到误差($\varepsilon=1.3\times10^{-6}$)要求。

如果要建立基于词的 Unigram 模型,则由于词的数量巨大,低频词的频率会更低,根据对 1995 年 2—7 月和 12 月的约 1600 万的《人民日报》语料的统计,二字以上的词出现频次在 15 次以上的词共有 21592 个;出现频次在 2～14 次的词共有 46637 个,这其中包括了不少的常用人名、地名以及数字;出现 1 次词共 9327 个,其中主要是一些人名和地名以及数字,当然也有一些和社会发展相适应的新词开始出现,比如"黑客"、"劝退"等。因此可以看出,低频词还是占据大多数,若 14 次以下的词都称作低频词,它们的频率约为 8.75×10^{-7},取 $p_0=8.75\times10^{-7}$,若仍然取可信度 $\delta=0.93,0.95,0.97,0.99$,则按式(20)计算得到表 2 所列出的结果。

表 2 基于词的 Unigram 模型训练样本规模

δ	0.93	0.95	0.97	0.99
$\Phi^{-1}(\frac{1+\delta}{2})$	1.81	1.96	2.17	2.58
N_w	3.38×10^7	3.96×10^7	4.84×10^7	6.84×10^7

若将 p_0 值取得更小,则要求的语料规模会进一步加大。由于表 2 考虑的建模语言单位是词,而根据对《现汉》的统计[8],二字词占总词数的 66.9%,一字词占 14.7%,三字词占 9.2%,四字词占 8.4%,五字以上的词占 0.8%。所以,其平均词长为 2.15,即使三字以上的长词出现较多,平均词长估计也不会超过 3。由此可见,建立以词为单位的 Unigram 模型所需训练语料规模最低估计为 $N=2.15\times N_w$。

如果考虑将式(6)中的绝对误差改为相对误差,即对其中的不等式 $\left|\frac{X}{N_w}-p\right|<\varepsilon$ 两边各除以 p,则式(6)变为:

$$P\left(\frac{\left|\frac{X}{N_w}-p\right|}{p}<\frac{\varepsilon}{p}\right)\geqslant\delta \tag{21}$$

由式(21)可知,使用频率越大的词,其经过训练以后的统计频率误差会越小。例如,如果在上式中,设可信度 $\delta=0.98$,$\varepsilon=4.375\times10^{-7}$,则对汉语中使用频率最高的"的"字,尽管在统计前无法确切地知道它的统计频率,但从已有的资料和统计中,能够粗略地估计出它的使用频率 $p>2.5\times10^{-2}$,将该统计频率和 δ、ε 的值代入式(21),就有 98%的把握保证"的"字统计频率的误差不会超过 $\varepsilon/p<4.375\times10^{-7}/(2.5\times10^{-2})=1.75\times10^{-5}$,而对那些使用频率较低的词,估计的相对误差就会大些。相对误差越小,利用 MLE 法所建立的语言模型的描述准确性就会越高。

结束语 由于 Unigram 模型是 n-gram 统计模型中的最简单一种,因此,它的训练语料样本规模可以看作是 n-gram 模型训练样本规模的下界,式(20)就是该下界的估计公式。它是考虑了建模语言单位在语言中的使用频率不同,其使用频率估计的误差要求就应该不同,得到的结果表明其更适于在实际中应用。

参 考 文 献

[1] 关毅. 基于统计的汉语语言模型研究[D]. 哈尔滨工业大学. 北京:国家图书馆,1999

[2] 刘爱芹. 随机抽样中样本容量确定的影响因素分析[J]. 山东财政学院学报,2006:60-64

[3] 张湘平,张金槐,谢红卫. 关于样本容量、验前信息与 Bayes 决策风险的若干讨论[J]. 电子学报,2003,31(4):536-538

[4] 王学民. 多指标分层抽样中样本容量折衷分配的加权方法[J]. 统计与决策,2008,3:27-29

[5] 耿修林. 整群抽样审计时样本容量的确定[J]. 审计研究,2007(6):85-88

[6] 盛骤,谢式千,潘承毅. 概率论与数理统计[M]. 北京:高等教育出版社,2001

[7] 北京语言学院语言研究所编. 现代汉语频率词典[M]. 北京:北京语言学院出版社,1986

[8] 刘小勤. 现代汉语分词词表的选词方法研究[D]. 太原:山西大学,1999