基于语料库的自然语言建模方法研究*)

张仰森[1,2] 曹元大[1]

(北京理工大学计算机科学与工程系 北京100081)[1] (山西大学计算机科学系 太原030006)[2]

摘 要 语言模型是对自然语言的描述,研究语言模型的构造方法是计算语言学的核心内容之一。本文在深入分析基于概率分布和上下文特征与信息的统计语言模型所依据的数学理论基础上,讨论了这两类语言模型的异同及其建立方法,并了讨论了基于组合思想的语言建模方法。

关键词 语言模型,语言建模,概率分布,上下文信息,组合思想

The Methods for Language Modeling Based on Corpus

ZHANG Yang-Sen[1,2] CAO Yuan-Da[1]

(Dept. of Computer Science & Engineering, Beijing Institute of Technology, Beijing 100081)[1]

(Dept. of Computer Science, Shanxi University, Taiyuan 030006)[2]

Abstract Language model is used to describe natural language, it is one of tasks of computational linguistics to study the ways of language modeling. This paper disscusses the model based on probability distribution and model based on context feature information, moreover, the difference and modeling methods of these two models have presented.

Keywords Language model, Language modeling, Probability distribution, Context information, Combination thought

语言模型是对自然语言的描述,研究语言模型的构造方法是计算语言学的核心。目前,语言模型主要包括规则语言模型和统计语言模型两大类。规则语言模型也称为基于知识的语言模型或唯理模型,它是根据语言学家的语言知识,利用形式语言和形式文法实现对语言的高度抽象描述。统计语言模型也称概率模型或经验模型,主要是利用数理统计方法,从大规模语料库中获取蕴含在其中的语言表达知识,并以这些知识构造表达语言的数学模型。十多年来,尽管基于大规模语料库的统计学方法在自然语言处理领域带来了成果,但这些方法似乎已发挥到了极致,要想取得新的、更大的、实质性的进展,还有待于从语言模型的建立方法或理论上实现重大突破。例如,将基于语言知识和逻辑推理的规则性方法与基于大规模真实语料库的统计性方法相结合可能就是一种方法上的尝试。本文试图通过对现有的一些语言模型结构及应用的分析,探讨语言模型的构造机理以及建立语言模型所依据的数学理论和所使用的方法,从而为人们在自然语言处理领域,针对不同的应用目标,选择适当的数学理论并使用适当的数学工具和方法,建造有效的、面向目标的语言模型指明方向。

1 基于概率分布的语言建模

基于概率分布的语言模型主要依据字词或字词对在文本中的分布概率或出现频率,利用概率统计理论和归纳方法,通过数学抽象,将概率或频次构成某种数学关系式,以实现对文本中字或词之间关系的表示,实现对文本的数学描述。由于不考虑过多的上下文信息,因而不涉及相邻词之间的转移概率问题,构造方法相对简单,一般可通过下列步骤实现概率分布模型的构造。

(1)统计计算文本中的词频或串频,以及词对出现的频率;

(2)利用最大似然法(MLE)等方法,求取词或词对的概率;

(3)利用概率统计理论和归纳学习方法,从统计数据中归纳抽提(或学习)出词语间的定性关系,形成表述词语间关系的语言模型;

(4)利用训练语料对模型中的参数进行定量估计和确定。

互信息模型就是基于概率分布的典型模型,它通过两个词的同现概率以及每个词在文本中的出现概率来反映文本中两个词间的联系强度,已被用在自然语言处理的许多领域。用下述的概率关系式表示词 w_1,w_2之间的相互关系:

$$MI(w_1,w_2)=\log_2\frac{P(w_1,w_2)}{P(w_1)*P(w_2)}$$

其中 $P(w_1,w_2)$是词对(w_1,w_2)的同现概率,$P(w_1)$,$P(w_2)$分别代表词 w_1,w_2的出现概率。当 w_1,w_2之间的联系关系较强时,$MI(w_1,w_2)>>0$;当 w_1与 w_2之间的联系较弱时,$MI(w_1,w_2)\approx 0$;而当 $MI(w_1,w_2)<0$时,w_1与 w_2在文本中的分布为互补分布。

当然,我们还可以根据应用的需要构造,利用概率分布构造新的模型。例如,在进行汉语分词时,为了解决歧义切分问题,可依据文本中字词或字词对的分布概率来构造用于分词歧义排除的模型如下。设 xyz 是一个有序的字串,如果 xy 和 yz 都是词,则在分词时,就会出现切分歧义。可定义下面的数学关系式解决字 y 与 x 或 z 成词的歧义问题[8]。

$$t_{x,z}(y)=\left[\frac{P(y,z)}{P(y)}-\frac{P(x,y)}{P(x)}\right]\div\sqrt{\frac{r(x,y)}{r^2(x)}+\frac{r(y,z)}{r^2(y)}}$$

其中 $P(x,y)$,$P(y,z)$分别是 x 与 y 以及 y 与 z 的同现概率;$r(x,y)$,$r(y,z)$分别表示 x 与 y 以及 y 与 z 的同现频次;$r(x)$,$r(y)$则分别表示 x 和 y 在文本中的出现频次。

这就是一个应用字或字对在文本中出现的频次与概率而

*)本文得到山西省青年科技研究基金的资助(20021015)。张仰森 博士生。曹元大 教授,博导。

构造的模型。由定义可知，当 $t_{x,z}(y)>0$时，y 有与后继 z 相连的趋势，值越大，相连趋势越强；当 $t_{x,z}(y)=0$时，不反映任何趋势；当 $t_{x,z}(y)<0$时，y 有与前趋 x 相连的趋势，值越小，相连趋势越强。它通过对三个汉字间结合力强弱的度量，实现歧义消除。

在汉语分词或文本校对中，为了判断两个相邻字的集合紧密程度或它们的接续关系，还可以利用以数理统计理论中的小概率原理构造一数学模型如下：设 Z_i 和 Z_{i+1}是两个相邻的汉字，现要判断它们间的接续关系。我们把 Z_i 和 Z_{i+1}独立作为结论假设 H_0，根据小概率事件原理对结论 H_0进行检验。选择一个数 $\alpha(0<\alpha<1)$作为显著性水平。Z_i 和 Z_{i+1}的接续关系由下式计算：

$$Rel(Z_i,Z_{i+1})=\frac{n\times(n_{11}\times n_{22}-n_{12}\times n_{22})^2}{n_{1\cdot}\times n_{2\cdot}\times n_{\cdot 1}\times n_{\cdot 2}}$$

其中，n 表示语料库中二元组的总数，n_{11}表示首字为 Z_i 次字为 Z_{i+1}的二元组的数量，n_{12}表示首字为 Z_i 次字不为 Z_{i+1}的二元组的数量，n_{21}表示首字不为 Z_i 次字为 Z_{i+1}的二元组的数量，n_{22}表示首字不为 Z_i 次字也不为 Z_{i+1}的二元组的数量。显然有 $n=n_{12}+n_{12}+n_{21}+n_{22}$，$n_{i\cdot}=n_{i1}+n_{i2}$，$n_{\cdot j}=n_{1j}+n_{2j}$，$(i,j=1,2)$。根据上面给出的建模步骤，求得模型的各个参数。$\alpha$ 的选择，可通过查上侧位表和实验相结合取得。当 $Rel(Z_i,Z_{i+1})>\alpha$ 时说明假设 H_0成立，Z_i,Z_{i+1}没有接续关系，否则，说明假设 H_0不成立，Z_i,Z_{i+1}有接续关系，$Rel(Z_i,Z_{i+1})$越小，接续强度越大。

2 基于上下文信息的语言建模

在任何一个真实的自然语言文本中，语言符号的出现概率是相互关联、相互制约的，只有考虑上下文之间的关系，才能真实地描述文本的本质，况且考虑的上下文信息越多，语言模型所反映的语序越逼近真实的语言句法模式。所以，这类语言模型也是近年来语言模型研究的重点，它在机器翻译、语音识别、文本校对等领域具有十分重要的应用价值。这类语言模型的建立，主要依据随机过程理论或信息论理论建立模型的数学结构，再通过统计方法和语句中的上下文信息实现模型参数的整定。

2.1 基于随机过程理论的模型构造

自然语言文本中各语言符号的出现概率并不相互独立，每个随机试验的结局依赖于它前面的随机试验的结局，也就是说，各语言符号的出现是相互联系相互制约的，它表明可以将自然语言文本看作一个随机过程。基于随机过程原理所建立的自然语言模型必能更加准确反映语言的本质，n-gram 模型和隐马尔可夫模型(HMM)多年来在自然语言处理领域长盛不衰的原因可能就在于此，因为 n-gram 模型和 HMM 实际上都是基于随机过程原理而建立的语言模型。

(1)n-gram 模型与 Markov 模型　Markov 模型是独立随机试验模型的直接推广，因1906年俄国著名数学家 Markov 对其研究而得名。Markov 模型指出，对一个随机过程$\{X_n,n\geqslant 0\}$，如果 $i_0,i_1,\cdots,i_{n-1},i,j$ 是一组不同的状态，它满足无后效性条件：$P\{X_{n+1}=j|X_n=i,X_{n-1}=i_{n-1},\cdots,X_0=i_0\}=P\{X_{n+1}=j|X_n=i\}$，则该随机过程 X_n 为离散时间的 Markov 链。随机过程有两层含义：第一，它是一个时间函数，随时间的改变而改变；第二，每个时刻上的函数值是不确定的，是按照一定的概率随机分布的。实际上，自然语言中每个字母或音素的出现随着时间的改变而改变，是时间的函数，而在每个时刻上出现什么字母(或音素)则有一定的概率性，是随机的。1913年，Markov 就注意到语言符号出现概率的相互影响，指出自然语言就是一个由有记忆信源发出的 Markov 链，在这一 Markov 链中，前面的语言符号对后面的语言符号是有影响的。

如果我们只考虑前面一个语言符号对后面一个语言符号出现概率的影响，这样得出的语言成分的链称作一阶马尔科夫链；如果考虑前面两个语言符号对后面一个语言符号出现概率的影响，则称作二阶马尔科夫链，以此类推，当考虑前面 n 个语言符号对后面一个语言符号出现概率的影响，则称作 n 阶马尔科夫链。随着马尔科夫链阶数的增大，随机试验所得出的语言符号链愈来愈接近有意义的语言文本。然而，正像语言学家乔姆斯基(Chomsky)所指出的，描述自然语言的马尔科夫链的阶数并不是无穷增加的，它的极限就是语法上和语义上成立的自然语言句子的集合，这样，就有理由将自然语言的句子看成是重数很大的马尔科夫链了。

n-gram 模型是近年来最流行的语言模型，它是这样定义的：如果用变量 S 代表文本中一个任意的符号(字、词、词性标记或义类标记符号)序列，它由顺序排列的 n 个符号组成，即 $S=W_1W_2\cdots W_i\cdots W_n$，则 S 在文本中的出现概率 $P(W_1W_2\cdots W_i\cdots W_n)$可以用下式表示：

$$P(S)=P(W_1W_2\cdots W_i\cdots W_n)=P(W_1)P(W_2/W_1)\cdots P(W_n/W_1W_2\cdots W_{n-1})$$

其中，$P(W_n/W_1W_2\cdots W_{n-1})$表示在给定上下文信息 $W_1W_2\cdots W_{n-1}$的条件下，W_n 的出现的概率，即要考虑前面的 $n-1$个符号对当前符号出现情况的影响。这种模型由于假设当前词的出现只与前面 $n-1$个词有关，而与其它词无关，可以看作满足 *Markov* 模型的无后效性条件，也就可以将其看作是一个广义的 $n-1$阶 Markov 模型。

(2)隐 Markov 模型　隐 Markov 模型是由 Baum 首先提出的，后被广泛地应用于语音识别和词性标注。它包含了双重随机过程，一个是系统状态变化的过程，状态变化所形成的状态序列叫做状态链；另一个是由状态决定观察的随机过程，是一个输出的过程，所得到的输出序列称作输出链。“隐”的意思就是输出链是可观察到的，但状态链却是“隐藏”的、看不见的。一个隐 Markov 模型的形式描述为 $\lambda=(A,B,\pi)$，其中，$A=\{a_{ij}\}$为状态转移概率矩阵，且$0\leqslant a_{ij}\leqslant 1$，$\sum_{j=1}^{N}a_{ij}=1$，$a_{ij}$表示从状态 i 转移到状态 j 的概率；$B=\{b_{ij}(K)\}$为输出概率矩阵，且$0\leqslant b_{ij}\leqslant 1$，$\sum_{j=1}^{N}b_{ij}=1$，$b_{ij}(K)$表示从状态 i 转移到状态 j 时输出为 K 的概率；π 为系统的初始概率分布。

(3)与基于概率分布的统计模型区别　由于考虑相邻词间的相互影响，基于随机过程理论所构造的语言模型中包含了相邻词间的转移概率，因此 Markov 和隐 Markov 模型中的一个很重要的参数就是状态转移矩阵，而在基于概率分布的统计模型中不会涉及词间的转移概率问题。

(4)建模时要考虑的因素　在构造 n-gram 模型时，主要涉及以下问题：a)模型参数 N 的选取。从理论上讲，N 值越大，所反映的语序越逼近真实的句法模式，因而会有更加良好的语法匹配效果。但在实际应用中，N 值的增大又会带来存储资源的急剧扩张和因统计数据稀疏而造成的计算误差；b)建模单元的选择。选择合适的建模单元，对模型的性能也有非常大的影响。例如，在对汉语言建模型时，可以字为单位建立模

型，也可以词为单位建立模型。当 N 相同时，基于词的模型优于基于字的模型，当然构造的难度也大，这是由于汉语中词的数量远多于字的数量。所以，要提高所建模型的性能，阶数 N 与建模单元的选择是需要权衡的一对矛盾；c)信道模型的选择。不同的信道模型适合于不同的应用对象，也可能导致不同的模型结构；d)状态转移矩阵的求取。这要通过对语料的统计处理才能得到。

在构建隐 Markov 模型时，首先是确定系统状态，并为系统的初始状态赋初值，形成系统的初始概率分布；然后，根据系统状态的数量以及语料统计数据获取状态转移矩阵以及输出概率矩阵，关键之处是参数的求取。

2.2 基于信息论最大熵方法的模型构造

最大熵方法是根据上下文建立语言模型的一种有效方法，它以上下文中对当前词的输出有影响的信息为特征，以信息论为理论依据，计算这些特征的信息熵，信息熵最大的特征说明对当前词的表征作用最强，并以这些特征构造模型。其问题描述如下：

设随机过程 P 所有的输出值构成有限集 Y，对于每个输出 $y\in Y$，其生产均受上下文信息 x 的影响和约束。已知与 y 有关的所有上下文信息组成的集合为 X，则模型的目标是：给定上下文 $x\in X$，计算输出为 $y\in Y$ 的条件概率，即对 $P(y|x)$ 进行估计。$P(y|x)$表示在上下文为 x 时，模型输出为 y 的条件概率。

由问题的描述可知，基于最大熵的建模方法涉及以下因素：

(1)特征。随机过程的输出与上下文信息 x 有关，但在建立语言模型时，如果考虑所有与 y 同现的上下文信息，则建立的语言模型会很繁琐，而且从语言学的知识上来讲，也不可能所有的上下文信息都与输出有关。所以在构造模型时，只要从上下文信息中选出与输出相关的信息即可，称这些对输出有用的信息为特征。

(2)特征的约束。与输出对象有关的上下文信息特征集合可能很大，但真正对模型有用的特征只是它的一个子集，因此，根据模型的要求对特征候选集中的特征进行约束。

(3)模型的构造和选择。利用符合要求的特征所构造的模型可能有很多个，而模型的目标是产生在约束集下具有最均匀分布的模型，而条件熵则是均匀分布的一种测量工具。最大熵的原理就是选择其中的一个条件熵最大的模型作为最后所构造的模型。

从上面的论述可知，基于上下文的语言模型主要依据随机过程的理论而建造，之所以应用 n-gram 模型和隐 Markov 模型，就是假设自然语言为具有 Markov 性的随机过程。那么，可否用 Poisson 过程或 Brown 运动来刻画自然语言呢？如果可以的话，是否就可建立具有新的结构的语言模型呢？当然也可以应用别的类似于最大熵的方法来构造语言模型。事实上，已有人将 Poisson 模型应用于文本检索，也有人尝试过 Brown 模型的应用。能够开阔人们的建模思路，也正是本文的初衷。

3 基于组合思想的语言建模

这种建模思想是将一些已有的语言模型通过数学插值、加权或剪枝嫁接等各种手段和方法，进行嫁接组合，形成新的描述力更强的语言模型。

3.1 利用插值法构造新的模型

该方法的思想是将不同语言模型通过插值的方法结合在一起，形成一个混合模型。假若要利用插值方法将阶数不同的语言模型结合在一起，形成新的模型，可令 $P^{(k)}(w_i|w_1^{i-1})$为第 k 个语言模型所决定的条件概率，则插值估计 $P^*(w_i|w_1^{i-1})$如下：

$$P^*(w_i|w_1^{i-1})=\sum_k \lambda_j(w_1^{i-1})P^{(k)}(w_i|w_1^{i-1})$$

例如，可以利用线性插值方法把 *Unigram*，*Bigram* 和 *Trigram* 结合在一起形成新的混合模型。

$$P^*(w_i|w_1^{i-1})=\lambda_1P(w_i)+\lambda_2P(w_i|w_{i-1})+\lambda_3P(w_i|w_{i-2}w_{i-1})$$

当然，插值方法可以是线性的，也可以是非线性的，插值系数由插值模型确定。这里介绍的是线性的，有关非线性的插值相对较为复杂，这里不再赘述。

3.2 利用最大熵法构造新的模型

Rosenfeld 利用最大熵原理探讨多知识源的模型集成方法，它将每个知识源视为一组限制条件的集合，加入到一个联合的估计之中。而这些限制条件的交集便是满足所有知识源的一组概率函数的集合，在这个集合中，具有最高熵值的函数，就是集成了所有知识源的概率函数。

3.3 利用加权方法构造新模型

加权方法是将一些模型进行加权处理，加权系数由人工经验或实验确定。例如，在计算中文文本中相邻汉字 Z_i 和 Z_{i+1}的同现概率时，为了防止由于语料规模问题而引起的某些二元组出现0次或很少次，可应用下面的加权模型计算字的同现概率：

$$P(Z_{i+1},Z_i)=\lambda_1F(Z_{i+1},Z_i)+\lambda_2F(Z_{i+1})$$
$$F(Z_{i+1},Z_i)=r(Z_i,Z_{i+1})/r(Z_{i+1})$$
$$F(Z_{i+1})=r(Z_{i+1})/N$$

其中，λ_1、λ_2为加权系数，$\lambda_1+\lambda_2=1$，可以根据人工经验或实验给出。$P(Z_{i+1},Z_i)$是 Z_{i+1}、Z_i 的同现概率，$r(Z_i,Z_{i+1})$为 Z_i，Z_{i+1}邻接同现次数，$r(Z_{i+1})$为 Z_{i+1}独立出现的次数，N 为汉语语料库中的字容量。

组合思想构造新模型的思路比较简单，实际中也被很多人采用，如何应用更好的组合算法，建立更加有效实用的模型还有待进一步研究。比如，如何将基于专家知识的规则模型与基于语料库的统计模型相结合构造新的有效模型，就值得研究。

4 语言建模的相关问题

4.1 语言模型参数的确定

在构建语言模型过程中，模型参数一般要利用训练语料才能确定。确定语言模型参数的过程可以看作是一个机器学习的过程，目前主要有两种方法：一种为有指导的参数求解，另一种为无指导的参数求解。有指导的参数求解是指有人工标注的熟语料库的情况下机器进行学习的过程，例如，在文[13]的汉语词性标注研究中，所选用的语言模型中涉及到两个参数，一个是二元语法项参数 $P(T_i|T_{i-1})$，它只与两个相邻词的词性有关，另一个是词典项参数 $P(W_i|T_i)$，它既与词性有关，又与词本身有关。只要从已标注好的训练语料集中，直接获取词 W_i 和它标记为词性 T_i 的次数 $R(W_i,T_i)$，以及词性 T_{i-1}与词性 T_i 的同现次数 $R(T_{i-1},T_i)$，就可以利用极大似然估计(*MLE*)法计算出模型的参数。所谓无指导的参数求解是指在没有熟语料的情况下的自学习过程。由于人工标注语料规模比较有限以及标注语料的不一致性(不同的人对同

一个词标注可能不同，甚至同一个人不同时间对同一个词的标注也不同）等弱点，无指导的参数估计方法正成为模型参数估计的主要方法。例如，Forward-Backward 算法[13]就是用来监督训练隐 Markov 模型的方法，利用它的前向算子和后向算子的反复迭代计算，可以得到隐 Markov 模型中的状态转移概率矩阵$\{a_{ij}\}$和$\{b_j(k)\}$。

4.2 数据稀疏问题的解决

在参数求解过程中，由于语料库规模的限制，可能会出现数据稀疏的问题。所谓数据稀疏就是指在整个训练语料中，有许多样本的出现概率很低甚至为0，它将会导致经过统计学习得到的参数是不可信的或是不充分的。目前解决数据稀疏问题的方法有各种数据平滑技术[10]以及词聚类法和词相似度法。词聚类方法的思想是通过将那些出现频率低的词抽象为词类，从而解决数据稀疏的问题，基于词类所构建的语言模型中，因为使用统计类的观察频率而不是词的观察频率，从而使参数空间大为缩小。词相似度法则基于如下的假设，即如果两个词 W_1与 W_2是相似的，那么对于一个参照词 W，互信息 $MI(W_1,W)$应接近于 $MI(W_2,W)$，$MI(W,W_1)$应接近于 $MI(W,W_2)$。然后根据词的相似度，来估计那些未出现的词对的互信息，解决数据稀疏的问题。有关各种数据平滑技术，读者可参阅文[10]。

4.3 语言模型的性能分析

语言模型的性能反映了模型描述自然语言的能力，对语言模型的评价是语言建模的一个很重要问题。通过语言模型评价，可以了解语言模型性能的优劣，从而指导模型的构建方法研究。针对各种语言模型建立有效的评价指标，是一个比较复杂和困难的问题，目前还没有一个好的解决办法，不过，针对最常用的 n-gram 模型，通常采用困惑度（Perplexity）来进行评价。困惑度的定义为如下：$Pe=2^H$。其中，$H=-\frac{1}{L-n+1}\sum_{i=n}^{N}\log_2 P(W_i|W_{i-n+1}^{i-1})$，$L$ 为测试样本的长度，n 为模型的元数。Pe 越小，说明该模型表述语言的能力越强。有关其它语言模型的评价方法还有待进一步研究。

结束语 本文在深入分析了目前广泛应用的自然语言处理模型的适应的领域、所依据的数学理论以及构建模型所采用的数学方法的基础上，澄清了多年来在语言模型理论方面人们的一些模糊认识，理清了有关语言模型理论上的一些思路，为人们针对不同的应用目标，依据适当的数学理论，使用适当的数学方法，建立更加准确的语言模型指明了方向。

语言模型是对自然语言进行各种处理的关键。多年来，尽管基于大规模语料库的方法为自然语言处理领域带来了成果，但从理论方法的角度考虑，有关计算语言模型的构建机理和方法还需进一步的研究。我们认为，有关语言模型理论的研究今后应在以下几个方面展开：

1. 构建语言模型的数学方法的研究。尽管本文对语言模型构造的方法进行了探讨，在分析现有的一些语言模型所依据的数学理论和所使用的方法的基础上，提出了一些建造语言模型的思想和方法，但这只能看作是抛砖引玉，希望引起人们研究语言模型数学理论的兴趣。

2. 语言模型结构参数的选择。这主要指建模单元和模型阶数的选择。例如，在中文处理中，应用 n-gram 模型，选择字或词作为建模单元以及不同的阶数 n，可得到性能不同的模型。

3. 语言模型的性能评价方法的研究。语言模型的性能或复杂度的评价问题，是指导建立语言模型的重要工具，但目前这方面的研究还比较少，针对 n-gram 的评价指标困惑度是根据经验人为建立的，并不具有普适性。如何从所构建的语言模型出发，应用严密的数学理论推导，探求出一套具有普适性的模型评价方法，从而给出比较严谨意义上的性能评价，还需进一步研究。

参 考 文 献

1 Rabiner L R. A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. Proc. IEEE，1989，77(2)：257～286

2 Rosenfeld R. A maximum entropy to adaptive statistical language learning. Computer Speech ang Language，1996，10(3)：187～228

3 Gao Jianfeng，Wang Haifeng，Li Mingjing，Lee Lai-Fu. A Unified Approach to Statistical Language Modeling for Chinese. Microsoft Research China Paper Collection，2000，1(9)：158～161

4 方兆本，缪柏其. 随机过程. 合肥：中国科学技术大学出版社，1993

5 复旦大学编. 概率论，第三册随机过程. 高等教育出版社，1981

6 傅祖芸. 应用信息论. 北京：电子工业出版社

7 李涓子，黄昌宁. 语言模型中一种改进的最大熵方法及其应用. 软件学报，10(3)：257～263

8 孙茂松，黄昌宁，等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展，34(5)：332～339

9 周强. 基于语料库和面向统计学的自然语言处理技术. 计算机科学，1995，22(4)：36～40

10 徐志明，王晓明，关毅. N-gram 语言模型的数据平滑技术. 计算机应用研究，1999(7)：37～39

11 孙茂松，黄昌宁，方捷. 汉语搭配定量分析初探. 中国语文，1997(1)：29～38

12 夏莹，常新功，等. 利用上下文相关信息的汉字文本识别. 中文信息学报，10(1)：23～30

13 白栓虎. 汉语词切分和词性标注一体化方法. 计算语言学进展和应用，清华大学出版社，1995

14 张树武，黄泰翼. 汉语统计语言模型的 N 值分析. 中文信息学报，12(1)：35～41