马尔克夫链与文本著作权自动判别（一）

许汉成

（黑龙江大学俄语语言文学研究中心，哈尔滨 150080；黑龙江大学计算语言学研究所，哈尔滨 150080）

提　要：本文介绍和分析了Д. В. Хмелев 采用的俄语文本著作权的定量分析方法。文本被看成是马尔克夫模型的实现。文本的著作权通过比较作者不明文本与已知作者文本集合的字母二元转移概率（一阶马尔克夫链）矩阵的距离而确定。通过著作权判别这一实例，文章展示了一个基于统计的自然语言处理系统建立的完整过程：建立统计模型，训练参数，测试模型，编码实现等。

关键词：马尔克夫链；N 元语法；统计模型；自然语言处理；著作权

中图分类号：H085.2　　**文献标识码：A**

一、前言

在文学和历史领域，著作权不明的现象不算罕见，比如围绕长篇小说《静静的顿河》的真实作者问题就出现了很多争议。产生著作权不明现象的原因很多。原稿遗失，匿名写作，冒名发表作品，剽窃，这些原因都可能造成著作权不明，甚至引起争议和司法程序。著作权是历史学、文学、语言学和法律证据学都十分关心的问题。

研究著作权的方法很多。文学家和语言学家在研究文本的著作权时一般会从文本的内容、思想、风格和语言的角度比较、分析争议的文本和可能的作者的典型文本。判定文本归属权时的主要语言性依据包括文字特点、语言风格、故事情节、特殊的词汇和语法结构，等等，但也有学者采用统计法。早在 19 世纪，Д. В. Морозов 就在这方面进行了探索。著名数学家 А. А. Марков 对 Д. В. Морозов 的方法提出了批评，认为正确的做法是自己统计《叶甫盖尼·奥涅金》的前 20000 个字母里元音和辅音的分布时所展示的方法（Марков 1913: 153-162; 1916: 239 – 242）。А. А. Марков 所说的正确方法就是马尔克夫链，现在它已经成为基于统计的自然语言处理系统的主要建模工具。В. П. Фоменко 和 Т. Г. Фоменко 选择了几个简单的作者风格参数，统计了大量 18—20 世纪俄罗斯作家的作品，认为虚词的比例是作家风格的常体（авторский вариант）（Фоменко 1983：86-109）。近年，Д. В. Хмелев 在著作权判定方面进行了新的探索，他利用马尔克夫链建立了著作权判定的数学模型，然后通过大量语料训练模型，经过测试，正确率达到了 70%以上，这里主要反映了他的研究成果（Хмелев 2000：115-126；Khmelev 2001：299-307；Кукушкина 等 2001）。

Д. В. Хмелев 的研究是一个十分典型的计算语言学应用实例，其中涉及语言的概率模型、统计建模型、马尔克夫链、N 元语法等概念，是学习基于统计的计算语言学思想和方法的良好教材。

Д. В. Хмелев 的方法的主要思想是：文本是马尔克夫模型的实现，通过一阶马尔克夫链建立作者已知文本的概率转移矩阵，用同样的方法计算建立作者不明文本的概率转移矩阵；

假定作者不明文本是由已知的几位作者之一创作的（不确切知道，到底是其中哪一位创作的），比较作者不明文本的转移概率分布与所有作者已知文本转移概率分布之间的距离，哪一个作者的文本转移概率分布与作者不明文本转移概率分布之间的距离最小，那么该作者是著作权不明文本作者的可能就越大（Хмелев 2000：115-126；Khmelev 2001：299-307）。

二、实验方案的数学基础

1. 马尔克夫链

马尔夫链是自然语言处理中使用最广的建模工具之一，那么什么是马尔克夫链呢？

简单地说，假定有一个符号串 $s=w_1, w_2, \ldots, w_n$，如果每个字符 w_i（$i = 1, \ldots, n$）的出现概率只跟前面出现的 j 个符号有关，那么这些符号的变化过程（或状态转移过程）就叫做 j 阶马尔克夫过程。（陈小荷 2000）这样的符号串就是马尔克夫链的一个实现。

最简单的情况是一阶马尔克夫过程，即每一个符号都只跟前面出现的一个符号相关，则符号串 $s=w_1, w_2, \ldots, w_n$ 的概率

$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2| w_1)\ldots P(w_n| w_{n-1})$

如果虚设 w_1，那么上面式子也可以表示为：

$$P(w_1, w_2, \ldots w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1})$$

符号 $\prod_{i=1}^{n}$ 表示求 n 项连续相乘的乘积。

类似地，在二阶马尔克夫过程中，一个符号的出现概率只跟它前面出现的两个符号相关，符号串的概率为：

$$P(w_1, w_2, \ldots w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1})$$

上面公式的实质是将整个符号串的概率转化为串里每个符号条件概率的乘积。

马尔克夫过程与 N 元语法很接近。N 元语法是指当前符号的条件概率取决于前面 N-1 个符号到它的转移概率。因此，三元语法相当于二阶马尔克夫过程，二元语法相当于一阶马尔克夫过程中。如果认为一个符号串的概率就是其中每个符号的概率的乘积，那就是一元语法。

现在举一个实际例子。假定我们有下面符号串 s：

```
GATCATTGATATGTTGCTAGAACTATGAGTGTTAAAGGTGCTTGTGGTGAGTTATCAGAC
AGAAACGCAGAAGATGTTATTGGAAGCTTGAGGAAAAGTGATCCTGGATTTACAGTGCCA
AGAATTGGCCTGTATTGTGTTCTCAATGTTTTTGAGGAAGGTAGAAACTGTAAGTGATGA
```

我们假定符号串 s 是由一阶马尔克夫过程生成的，也就是说，每个符号的出现概率只跟它前面出现的一个符号相关。现在要知道生成该符号串的马尔克夫模型。

经过观察、统计，我们发现这个符号串只是由 4 个不同的字母组成的：A、C、G、T，其中 A 出现 54 次，C 出现 19 次，G 出现 51 次，T 出现 56 次，总长度为 180 个字母。

为了突出开始和结尾，我们在上面符号的前后各添加一个特殊的非字母符号，如“_”，这样便得到下面符号串 S′：

_GATCATTGATATGTTGCTAGAACTATGAGTGTTAAAGGTGCTTGTGGTGAGTTATCAGAC
AGAAACGCAGAAGATGTTATTGGAAGCTTGAGGAAAAGTGATCCTGGATTTACAGTGCCA
AGAATTGGCCTGTATTGTGTTCTCAATGTTTTTGAGGAAGGTAGAAACTGTAAGTGATGA_

假定有个小窗子，宽度为 2 个字母，我们拿着这个的想象小窗子，从左向右一个字母一个字母地扫过符号串 s′。我们透过小窗子每次可以看到两个字母，累计不同字母对的出现次数，可以得到下面表格：

S′的二元及其频数

第一个字母	第二字母					
		A	C	G	T	_
	A	17	5	17	14	1
	C	7	3	1	8	
	G	20	6	8	17	
	T	10	5	24	17	
	_			1		

这个表格说明了每个 A、C、G、T 四个字符(加上辅助符号“_”)组成的所有二元(bigrams)及其在 S′里的出现次数。为了更加直观，我们可以画出一个状态转移图。S 里的二元与 S′基本一致，只是要去掉“_G”和“A_”。

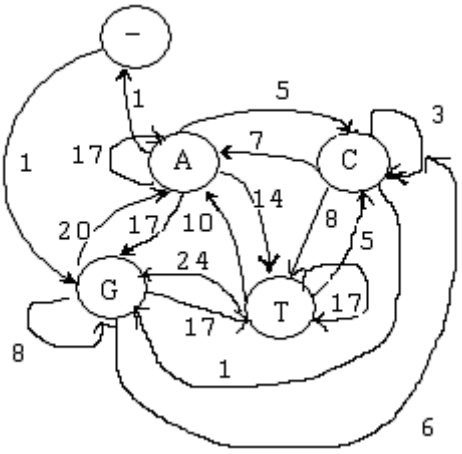

图 1. S’ 的状态转移图

有了上面基本数据后，我们可以计算二元的转移概率矩阵。

	第二个字母				
第一个字母		A	C	G	T
	A	0.320755	0.094340	0.320755	0.264151
	C	0.368421	0.157895	0.052632	0.421053
	G	0.392157	0.117647	0.156863	0.333333
	T	0.178571	0.089286	0.428571	0.303571

N 元的频数和频率、转移概率这些数据是比较文本的基本依据。当为已知类型的文本建立状态转移概率模型之后，遇到类似未知的新文本，同样为新文本建立相应的概率模型，然后就可以比较新文本与已知文本的概率模型的距离。计算距离的方法很多，比如相对熵。

2. 文本著作权自动识别的数学模型

假定 A 为字母的有限集合。A^r 表示定义在字母表 A 上的长度为 r 的字符串，A 上所有长度（按字母个数计算）的字符串可以表示为 $A^*=\cup_{r>0} A^r$（∪ 是集合的并运算符）。字符串 $f\in A^*$ 的长度表示为 $|f|$。

识别作品著作权的任务可以表述为：假定有 W 个作者，每个作者有 N_w 个文本；每个作者构成一个类型，用 C_w 表示，$w=0,\ldots W-1$，每个作者类型 C_w 的每个文本片断表示为 $f_{w,n}$，那么 $f_{w,n}\in A^*$，$n=1,\ldots,N_w$。因此 C_w 可以看成是 $f_{w,n}$ 的集合，即 $C_w=\{f_{w,n}|n=1,\ldots,N_w\}$。我们的任务是把未知文本 $x\in A^*$ 归入 C_w 之一。

假定文本片断 $f_{w,n}$ 是马尔克夫链的实现。按照上一节介绍的方法，建立每位作者的任意一对字母的转移概率矩阵 $P^w(w=0,\ldots W-1)$。用 $h_{w,n,kl}$ 表示文本片断 $f_{w,n}$ 里字母对 k->l 的转移频数，那么作者 C_w 的所有已知作品里的 k->l 的转移频数为 $h_{w,kl}=\Sigma_n h_{w,n,kl}$，而作者 C_w 的所有已知作品里 k 的频数为 $h_{w,k}=\Sigma_l h_{w,kl}$，根据条件概率的定义，$P_{kl}^w=h_{w,kl}/h_{w,k}$。某些字母对 k->l 的转移概率可能为 0，建立有序对（k,l）集合 Z_i，并使 $P_{kl}^w>0$。

假定有作者未明文本 x，它是转移概率矩阵为 P^θ 马尔克夫链的实现，其中 θ 为未知参数，取值范围为 $0,\ldots,W-1$。

现在用 $v_{k,l}$ 表示 x 里的 k->l 转移频数，同样 $v_k=\sum_l v_{k,l}$，令

$$L_w(x)=-\sum_{(k,l)} v_{k,l}\times\ln(v_{k,l}/(P_{kl}^w\times v_k)),\quad ①$$

有序对（k,l）全部从 Z_i 里取。这个公式将 x 里的有序对 $v_{k,l}$ 与已知的每个作者的概率转移矩阵进行比较，也就是求 x 与作者 w 之间的距离。其实，$P_{kl}^x=v_{k,l}/v_k$，因此上面公式也可以改写为

$$L_w(x)=-\sum_{(k,l)} v_{k,l}\times\ln(P_{kl}^x/P_{kl}^w),$$

经过改写，公式①就更像相对熵的公式了：

$$D(p\|q)=-\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)}$$

相对熵计算的是两个概率分布 P(x)与 q(x)之间的距离。Д. В. Хмелев 比较未知作者与已知作者距离的公式实际上与相对熵的计算方法是一致的。

很明显，在求出 x 与所有作者的距离之后，对这些距离值进行比较，可以得到 W 个距离值，即

$$\{L_w(x)\,|\,w=0,\ldots,W-1\}$$

对 $L_w(x)$按照从小到大的方向排序，最小的序号为 0，依次递增。

与 x 距离最短的作者应该最有可能是作者不明文本的真实作者，就是说序号为 0 的那个

作者的编号（w,…W-1）就是 θ 的估计值。而且，序号越低，风格越接近文本 x。假定 t(x) 是参数 θ 的最大似然估计，那么

$$t(x) = \arg\min_{w=0,\ldots,W-1} L_w(x) \quad ②$$

这样，我们最终得到参数 θ 的估计值，这个值表示作者不明作品的可能真实作者序号。显然， θ 的取值范围必须在已知作者集合里，因为不可能求 x 与未知作者的距离。

三、实验方案与文本预处理

取集合 A = {小写西里尔字母}∪｛空格符｝。假定我们有 W 个作者用俄语创作的若干个足够长的文本片断。用 $g_{w,n}$ 表示第 w 个作者的第 n 个文本片断。那么，可以认为 $g_{w,n}$ 是由字母表 B 里的字母序列，字母表 B 是字母集合 A 的扩展，除 A 里规定的符号外，还包括各种大写字母、标点符号、拉丁字母等（在个人计算机上 B 通常与扩展 ASCII 码表一致）。

每个文本片断 $g_{w,n}$ ∈B*可以通过函数映射到 A*，即 F: B* —> A*。这个函数描述的实际上是一个文本预处理过程。实验表明，删除所有大写字母开始的词（包括句首大写的词）可以显著提高系统判定著作权的成功率，库古什金娜解释说，文学作品人物姓名与作者风格没有关系，我们觉得可能是因为文学作品大量、反复出现的人物姓名影响二元频率（Кукушкина 等 2001）。在进行预处理时，字母 ё 也并入 е，这样，加上空格共 33 个不同符号，每个字母都编上号：字母 а 对应于 1，…，字母 я 对应于 32，空格的代码为 33。理论上，由 33 个字符组成的二元有 33×33 = 1089 个。实际上，由于一些字母转移概率为 0，实际字母二元的个数有 1000 个左右。

举个预处理的例子，假如有字符串 y = “Кроме того, мы будем рассматривать функцию G”，那么

x = G(y) = “того мы будем рассматривать функцию ”

x 就是 y 的预处理结果。

现在假定 y∈B*是 W 个作者之一，但是我们不能确定，到底属于哪一位作者。我们现在的任务就是判断文本片断 y 的作者。我们将公式①用于 x = G(y)，然后就可以根据公式②估计文本片断 y 的真实作者。字母二元的判定著作权方法没有利用俄语词法、句法等语言知识。语言学家或许会怀疑这种方法是否可行，但是，Д. В. Хмелев 经过实验证明，利用字母二元判定著作权方法还是比较可靠的，成功率达到 70%以上。　　（未完待续）

The Identification of Text Authorship Using Markov Chain(I)

XU Han-cheng

(Centre for Russian Language and Literature Studies of Heilongjiang University, Harbin, 150080, China; Research Institute for Computer Linguistics of Heilongjiang University, Harbin, 150080, China)

Abstract: The paper introduces and analyses a technique for text authorship identification used by D. V. Khmelev. Text is viewed as the realization of letter bigrams (first-order Markov chain). The authorship is determined by the measurement between distances of transitional probability matrices of text with

unknown authorship and sets of texts from known author. The paper exemplifies the whole process of statistical natural language processing: statistical modeling, parameters training, test and coding.

key words: Markov chain; N-gram; statistical modeling; NLP; authorship identification

收稿日期: 2009-10-28

作者简介: 许汉成（1965－），男，陕西汉中人，解放军国际关系学院副教授，博士，黑龙江大学俄语语言文学研究中心兼职研究员。研究方向：俄语语言学，计算语言学。

[责任编辑：叶其松]