马尔克夫链与文本著作权自动判别（二）

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. Шмелев 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**

四、初步实验

初步实验主要检验模型的正确性。Д. В. Хмелев 首先选择了的 4 位作家：К. Булычев, А. Волков, Н. В. Гоголь 和 В. Набков。

实验实际上是检验 t(G(y))的有效性。具体方法如下：将每位作家 w（w = 0, 1, 2, 3）的作品分为两部分，保留一篇作为测试作品，记作 y^j (j = 0,1,2,3)，其它作品 $f_{w,n}$ 用来训练模型参数，获得字母二元转移概率矩阵 P^w，然后求 t (F(y^j)) 。如果模型工作正常，那么 t (F(y^j)) = j。

初步实验文本集合如下：

0）К. Булычев: Умение кидать мяч (y^0); Белое платье золушки ($g_{0,1}$); Великий дух и беглецы ($g_{0,2}$); Глубокоуважаемый микроб ($g_{0,3}$); Закон для дракона ($g_{0,4}$); Любимец [Спонсоры] ($g_{0,5}$); Марсианское зелье ($g_{0,6}$); Миниатюры ($g_{0,7}$); “Можно попросить Нину?” ($g_{0,8}$); На днях землетрясение в Лигоне ($g_{0,9}$); Перевал ($g_{0,10}$); Показания Оли Н. ($g_{0,11}$); Поминальник XX векаа ($g_{0,12}$); Раскопки курганов в долине Репеделкинок ($g_{0,13}$); Тринадцать лет пути ($g_{0,14}$); Смерть этажом ниже ($g_{0,15}$);

1) А. Волков: Семь подземных королей (y^1); Волшебник изумрудного города ($g_{1,1}$); Урфин Джюс и его деревянные солдаты ($g_{1,2}$); Огненный бог Марранов ($g_{1,3}$); Гениальный пень ($g_{1,4}$); На войне, как на войне ($g_{1,5}$); О чем молчали газеты ($g_{1,6}$); Преступник и наказание ($g_{1,7}$); Эпилог ($g_{1,8}$); Желтый туман ($g_{1,9}$); Тайна заброшенного замка ($g_{1,10}$);

2)Н.В. Гоголь: Рассказы и повести (y^2); Ревизор ($g_{2,1}$); Тарас Бульба ($g_{2,2}$); Вечера на хуторе близ Диканьки ($g_{2,3}$);

3) В. Набоков: Другие берега (y^3); Король, дама, валет ($g_{3,1}$); Лолита ($g_{3,2}$); Машенька ($g_{3,3}$); Рассказы ($g_{3,4}$); Незавершенный роман ($g_{3,5}$);

下来的任务就是利用训练文本统计每一个初步实验作家的字母二元转移概率矩阵，即该作

家的概率模型 P^w(这里 w=0,1,2,3)，然后计算每个测试文本的字母二元转移概率矩阵，接着计算每个测试文本与每位已知作家文本的距离。根据 Д.В. Шмелев 的计算， 初步实验的结果如下：

作者（P^w） $L_w(y^j)$ 测试文本 (y^j)	К. Булычев （P^0）	А. Волков （P^1）	Н.В. Гоголь （P^2）	В. Набоков （P^3）
y^0	2.484569	2.508425	2.504301	2.49377
y^1	2.501061	2.473907	2.516797	2.492874
y^2	2.499033	2.504508	2.480202	2.483928
y^3	2.541367	2.538101	2.548842	2.520018

上面表格也可以用 $y^j \times P^w$ 的矩阵表示：

$$L = \begin{pmatrix} 2.484569 & 2.508425 & 2.504301 & 2.49377 \\ 2.501061 & 2.473907 & 2.516797 & 2.492874 \\ 2.499033 & 2.504508 & 2.480202 & 2.483928 \\ 2.541367 & 2.538101 & 2.548842 & 2.520018 \end{pmatrix}$$

仔细观察上表，每一个测试文本与四位潜在的作者都有一个距离值，如测试文本 y^0 与四个潜在作家的距离分别是 $L_0(y^0)=2.484569$，$L_1(y^0)=2.508425$，$L_2(y^0)=2.504301$，$L_3(y^0)=2.49377$，所以 $L_0 \leqslant L_3 \leqslant L_2 \leqslant L_1$。我们将 L_0、L_1、L_2 和 L_3 按照由小到大的方向排序，最小的顺号为 0，最大的序号为 3，因此 L_0、L_1、L_2 和 L_3 的序号分别是：0、3、2、1。以此类推，我们可以得测试文本为 y^1、y^2、y^3 时 L_0、L_1、L_2 和 L_3 的序号，于是得到下面 y^j x L_w 矩阵：

$$R = \begin{array}{c} \begin{matrix} L0 & L1 & L2 & L3 \end{matrix} \\ \begin{pmatrix} 0 & 3 & 2 & 1 \\ 2 & 0 & 3 & 1 \\ 2 & 3 & 0 & 1 \\ 2 & 1 & 3 & 0 \end{pmatrix} \end{array} \begin{matrix} y0 \\ y1 \\ y2 \\ y3 \end{matrix}$$

大家还记得，每个作家的作品分为两个部分，$L_j(y^j)$计算的是 y^j 与作家 w = j 之间距离（更准确地说，是二者转移概率分布之间的距离）。所以，$L_j(y^j)$的序号应该是几个距离值$\{L_w(y^j) \mid w=0,1,2,3\}$中最小的，即 $L_j(y^j)$的序号为 0 说明模型正确。从矩阵 R 来看，就是 R 的主对角线上的值全部为 0。如果 R 主对角线上的值不为 0，就说明模型判定作者出错了，值是几，说明与正确序号的误差是几。

实验结果如下：

N	作者	序号 （c1）	文件数 （c2）	训练文本 总字符数 （c3）	测试文本总 字符数 （c4）
0	К. Булычев	0	15	2345689	75161
1	А. Волков	0	8	1733165	233418
2	Н.В. Гоголь	0	3	723812	243767
3	В. Набоков	0	5	1658526	367179

c2 栏是文件总数（作品与文件不完全一致），c3 为训练文本的字符总数（$c3 = |F(g_{w,n})|$），c4 为测试文本总字符数（$c4 = |F(y^j)|$）。对于 К. Булычев 来说，训练文本的字符总数 c3=Σ

${}_nF(g_{0,n})$＝2345689，而测试文本字符总数（即"Умение кидать мяч"）c4=｜F (y^0)|＝75161。c1 为 $L_j(y^j)$在数列$\{L_w(y^j) \mid w=0,1,2,3\}$中的序号，它们的值都为 0，说明 4 个测试文本的著作权判定结果全部正确。

五、大量文本试验

初步实验时，参与实验的作者仅有 4 人，文本仅 33 个，模型的正确性必须通过更大规模的实验进行验证。大规模实验使用 $W = 82$ 个作者的作品，每位作者的文本数量从 1 到 30 个不等。如果只找到某位作者的一个文本，这个文本就要分成两个部分，分别用作训练文本和测试文本。根据经验，当文本字符数超过 100000 个模型的性能比较好，所以在选择作者时，只选择文本规模超过上面最低要求的作者。作品的总数超过 1000 个，它们被分成了 386 个文件，文本字符总数超过 128×10^6。

这样，矩阵 L 和 R 都是 82×82 方阵。

实验过程中，为每位作者建立了作品 $g_{w,n}$，从中计算出 P^w，同时保留 1 个文本作为测试文本 y^j。实验结果表明，正确答案（c1=0）的数目很高，达到 69，真实作者在可能作者列表里排在第 2 的情况有 3 次（c1=1），排在第 3 的情况有 2 次（c1=2），排在第 4 的情况有 1 次（c1=3）。其它 7 位作者文本的判定错误较大，真实作者被排在了 10 名以外。

作者	1 2	c_3	c_4
К. Булычев	5	200 7724	64 741
О. Авраменко		173 3113	22 3718
А. Больных		129 4721	37 3611
А. Волков		147 8932	20 2495
Г. Глазов		139 8323	18 4593
М. и С. Дяченко		175 4213	19 7039
А. Етоев		267 096	80 358
А. Кабаков		905 502	22 2278
В. Каплан		515 029	12 9608
С. Казменко		184 6161	15 6768
В. Климов		250 231	17 9903
И. Крашевский		118	48

0

1				3722	1795
2	И. Кублицкая			282377	170469
3	Л. Кудрявцев			583239	179093
4	А. Курков			628041	218726
5	Ю. Латынина	0		2628781	283565
6	А. Лазаревич	6		310553	94629
7	А. Лазарчук			2395669	210151
8	С. Лем			1568013	343519
9	Н. Леонов			568854	279377
0	С. Логинов	4	3	1998543	159247
1	Е. Лукин			602216	125694
2	В. Черняк			920056	201636
3	А.П. Чехов			662801	343694
4	И. Хмелевская			1524905	203684
5	Л. и Е. Лукины			837198	122999
6	С. Лукьяненко		4	3682298	483503
7	Н. Маркина			266297	93647
8	М. Наумова			306514	337821
9	С. Павлов			751836	453448
0	Б. Райнов			1405994	420256
	Н. Рерих			101	21

1			1285	1047
2	Н. Романецкий		130 5096	11 7147
3	А. Ромашов		884 34	87 744
4	В. Рыбаков		715 406	12 1497
5	К. Серафимов		186 424	75 276
6	И. Сергиевская		109 118	50 786
7	С. Щеглов	0	253 732	55 188
8	А. Щеголев		848 730	10 5577
9	В. Шинкарев	9	156 667	80 405
0	К. Ситников		419 872	10 9116
1	С. Снегов		824 423	40 8984
2	А. Степанов		122 3980	93 707
3	А. Столяров	1	350 053	13 7135
4	Р. Светлов		454 638	26 8472
5	А. Свиридов	3	660 413	23 5439
6	Е. Тильман		705 352	46 4685
7	Д. Трускиновская		200 5238	11 8351
8	А. Тюрин	8	410 9050	11 0237
9	В. Югов		829 209	66 657
0	А. Молчанов		398 487	20 6541
	Ф.М. Достоевский		613	88

1		825	582
2	Н.В. Гоголь	638339	215540
3	Д. Хармс	199449	114889
4	А. Житинский	2137325	543037
5	Е. Хаецкая	723167	204091
6	В. Хлумов	788562	183358
7	В. Кунин	1335918	296463
8	А. Мелихов	615548	458086
9	В. Набоков	1522633	342774
0	Ю. Никитин	1342176	702383
1	В. Сегаль	320218	75917
2	В. Ян	507502	600636
3	А. Толстой	129664	97842
4	И. Ефремов	536604	256521
5	Е. Федоров	1120665	221388
6	О. Гриневский	158762	96085
7	Н. Гумилев	70181	71042
8	Л.Н. Толстой	1225242	199903
9	В. Михайлов	254464	84135
0	Ю. Нестеренко	352988	71075
	А.С. Пушкин	170	57

1			380	143
2	Л. Резник		115 925	79 628
3	М.Е. Салтыков-Щедрин		239 289	10 1845
4	В. Шукшин		309 524	66 756
5	С. М. Соловьев		234 5807	16 0002
6	А. Кац		841 898	81 830
7	Е. Козловский		849 038	88 9560
8	С. Есенин		219 208	44 855
9	А. Стругацкий		151 246	51 930
0	А. и Б. Стругацкие	9	657 1689	34 5582
1	Б. Стругацкий		298 832	26 1206

六、结论

Д.В. Шмелев 等人的研究表明，以字母为单位的一阶马尔克夫链是一种有效的文本著作权建模方法。这种方法简单易行，不需要深层次的语言自动分析，不仅适用俄语文本，同时也可以推广到其它拼音文字文本。利用语法信息有助于判定文本的著作权。但是，需要注意的是，使用大量语言范畴和提高分析深度不一定能够提高系统的精度。这是一个应该进一步探讨和研究的问题。

二个连续字母反映了文本的词形的结构（前缀、词根、后缀和词尾），包含了一定作者语言的词汇、语法使用的特点，这可能是字母二元法在著作权判定方面取得成功的主要原因。

基于统计的自然语言处理系统容易实现、功能强大，是大规模真实文本处理的主要方法（黄昌宁 2002）。Д. В. Шмелев 的研究展示了基本统计的自然语言处理系统建立的全过程：明确问题的已知条件和研究目标，建立一个统计语言模型，接着确定模型的训练和测试文本集，通过实验训练模型，验证模型的正确性，最后建立语言处理程序[1]。我们在中外文处理问题的研究中也应该掌握并拿起这一强大武器。

附注

1 该本著作权判别系统可网络访问、测试（http://rusf.ru/books/analysis/about.htm）

参考文献

[1]Khmelev Dmitri V. and Tweedie Fiona J. 2001 Using Markov Chains for Identification of Writers[J]. Literary and Linguistic Computing, , Vol.16, №.4.

[2]Кукушкина О. В., Поликарпов А. А. ,Хмелёв Д. В. 2001 Определение авторства текста с использованием буквенной и грамматической информации. «Проблемы передачи информации»[J]. Т. 37, Вып. 2. Available online: http://www.philol.msu.ru/~lex/articles/grco_r.htm.

[3]Марков А. А. 1916 Об одном применении статического метода[J]. Изв. Имп. акад. наук., №4.

[4]Марков А. А. 1913 Пример статического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь[J]. Изв. Имп. акад. наук., №3.

[5]Морозов Н. А. 1915 Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд[J]. Известия отд. русского языка и словесности Имп. акад. наук., Т.20, Кн.4.

[6]Фоменко В. П., Фоменко Т. Г. 1983 Авторский инвариант русских литературных текстов[J]. Метод количественного анализа текстов нарративных источников. М., Ин-т истории СССР.

[7]Хмелев Д. В. 2000 Распознавание автора текста с использованием цепей Маркова[J]. Вестн. МГУ. Сер. 9, Филология.. №2.

[8]陈小荷 2000 现代汉语自动分析——Visual C++实现[M]。北京语言文化大学出版社。

[9]黄昌宁 2002 统计语言模型能做什么？，语言文字应用，第 1 期。

The Identification of Text Authorship Using Markov Chain (Ⅱ)

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: 2009-06-28

作者简介: 许汉成（1965－），男，陕西汉中人，俄语副教授，博士。研究方向：俄语语言学，计算语言学。

[责任编辑：叶其松]