中国人名性别自动识别

郎君，秦兵，刘挺，李生

（哈尔滨工业大学信息检索研究室，哈尔滨 150001）

摘 要： 人名性别识别能应用在自然语言处理和信息检索中。本文尝试了中国人名性别自动识别的两种方法。一种方法是采用贝叶斯方法对比了三种人名用字模型，对 10 万人名的实验结果表明，人名尾字对性别识别具有更好的应用能力，开放测试准确率为 82.95%。另一种方法依赖人名上下文，从 Hownet 和网络挖掘分别抽取男、女性别指示词，采用在百度检索人名的结果中对性别指示词计数来获得对应的性别。对 81 个人名的测试，准确率达到 96.3%。结果显示 Hownet 的性别指示词具有较好的通用性，网络挖掘的性别指示词具有较好的领域适应性。

关键字： 性别识别；贝叶斯方法；性别指示词；网络挖掘

Gender Recognition of Chinese Person Name

Jun Lang, Bing Qin, Ting Liu, Sheng Li

(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001)

Abstract: Gender recognition of person name can be used in natural language processing and information retrieval. This paper has tried two methods on automatical gender recognition of Chinese person names. The one used bayes statistics for comparing three models of Chinese person name characters. On 100,000 person names, experiment showes the end character has better application ability. The open test accuracy achieved to 82.95%. The other method used context for gender recognition. Hownet and web mining were applied for extracting gender designators. In the returning snippets from Baidu, male and female gender designators were counted for judging the final gender type. On 81 person names, the best accuracy in context based experiment was 96.3%. Experiment showes the gender designatores from Hownet have better universality and that from web mining have better adaption to fields.

Keywords: Gender recognition; Bayes method; Gender designator; Web mining

1 引言

随着自然语言处理和信息检索研究的不断深入，各种深层的语言理解技术不断提上日程，人名性别识别就是其中之一。人名性别识别根据人名的用字特点或所在上下文识别出当前人名的性别，可以用在指代消解[1]、机器翻译[2]、检索结果后聚类[3]中。

英文上，人名性别识别最早是将常见男、女名存放数据库中，确定人名性别时，在数据库中检索[4]。句法分析器 Minipar 采用了类似方法[5]。英文中结合上下文确定人名性别主要采用性别模板的方法。文献[6]中采用了两类性别模板：基于 Minipar 分析结果的模板，和 web 检索结果的性别指示模板。中文方面，文献[7]对 7 万中国人名的 90 个常用尾字进行非参数检验后认为：男女人名用字有显著性别差异。文献[8]、[9]都指出男性和女性对于

基金资助：国家自然科学基金 60435020, 60575042, 60503072; 腾讯基金项目

作者简介：郎君（1981－），男，四川峨眉人，哈工大计算机系博士研究生，bill_lang@ir-lab.org

语言的态度不一样，在描写男性和女性的时候，它们的上下文存在不同的性别语义场。

本文组织如下：第二部分介绍依赖人名用字的性别识别，采用朴素贝叶斯方法，详细考察了中国人名中间字、尾字和性别的关系；第三部分提出依赖上下文的人名性别识别，主要是根据人名上下文词汇的整体性别指示性来确定人名性别；第四部分是结论和展望。

2 依赖人名用字的性别识别

2.1 建立模型

对每个中国人名 *name*，取 $name = w_0 w_1 w_2$。w_0 为人名的姓氏，如张，王，慕容[1]等；w_1 为中间字，对于二字的人名，或者是复姓的三字人名，w_1 为空格；w_2 为人名尾字。采用 Naïve Bayes 统计方法来预测每个中国人名对应的性别。

$$gender^* = \arg\max_{gender \in [m,f]} p(gender \mid name) = \arg\max_{gender \in [m,f]} \frac{p(name \mid gender)p(gender)}{p(name)} \quad (1)$$
$$= \arg\max_{gender \in [m,f]} p(name \mid gender)p(gender)$$

公式(1)中 *gender* 表示可能的性别，*m* 表示男性，*f* 表示女性，$gender^*$ 表示 *name* 对应最可能的性别。对于中国人名用字和性别的关系，可以有如表 1 所示的三种模型。

表 1 人名用字和性别的三种 Naive Bayes 模型

Tab.1 Three Navie Bayes Models of Gender and Characters of person name

编号	假设	公式
1	w_2	$gender^* = \arg\max_{gender \in [m,f]} p(w_2 \mid gender)p(gender)$
2	$w_1 w_2$（独立）	$gender^* = \arg\max_{gender \in [m,f]} p(w_1 \mid gender)p(w_2 \mid gender)p(gender)$
3	$w_1 w_2$（整体）	$gender^* = \arg\max_{gender \in [m,f]} p(w_1 w_2 \mid gender)p(gender)$

模型 1 假设性别只和中国人名尾字有关。模型 2 假设人名的中间字和尾字没有任何关系，对于一些三字人名考虑中国传统家族排行的因素，中间字和尾字是相互独立的。模型 3 假设性别同时和中国人名的中间字和尾字有关，比如“秀丽”、“雅丽”、“美霞”等。我们把中间汉字和尾字作为一个整体来进行统计分析。

2.2 实验和结果分析

在文献[1] 的 4 万人名基础上，加入 6 万人名，按照 3：1 随机分割，得到表 2 所示的实验数据。这 10 万人名对应的是目前中国高校 18 岁到 30 岁之间的年轻人。按照 2.1 的模型准备，得到表 3 所示的实验结果。

表 2 实验相关的 10 万人名分散情况

Tab.2 The distribuation of 100,000 experimental person names

性别	训练数据(3/4)	测试数据(1/4)
男性	56481	18827
女性	24786	8263

[1] 二字的姓是复姓，中国人名中常见的复姓有慕容、司马、公孙等。

表 3　10 万人名上的贝叶斯试验结果

Tab.3 Bayes experimenal result on 100.000 person names

编号	假设	测试类型	男性人名			女性人名			整体准确率
			识别正确	识别错误	准确率	识别正确	识别错误	准确率	
1	w_2	封闭	53615	2866	94.93%	16672	8115	0.672611	86.49%
		开放	17279	1548	**91.78%**	5191	3071	**0.628298**	**82.95%**
2	w_1w_2（独立）	封闭	47005	9476	83.22%	15658	9129	0.631702	77.11%
		开放	14675	4152	77.95%	5488	2774	0.664246	74.43%
3	w_1w_2（整体）	封闭	54065	2416	**95.72%**	21260	3527	**0.857708**	**92.69%**
		开放	16194	2633	**86.01%**	5587	2675	**0.676229**	80.41%

表 3 显示，在开放和封闭测试中，模型 2 效果都最差。说明将人名的中间字和尾字考虑成完全相互独立是不恰当的。同时也说明，现在中国的年轻一代的名字已经较少考虑中国传统的家族排行了。

模型 3 在封闭测试中，各种准确率都同比最好。说明完全可见的样本仅使用尾字不能达到最好效果，中间字对性别具有很好的指示，而且这种指示性需要结合尾字来发挥。

开放测试中模型 1 总体效果最好。模型 3 封闭测试上效果较好，但开放测试时比模型 1 差一些。假设中文有 6 千常用字，模型 3 可能遇到的情况有 360 万，而实验仅有几万数据规模，出现了严重的数据稀疏问题。模型 1 可能的情况只有 6000 个，实验的几万数据可以覆盖多数情况。所以出现上述实验结果。

通过现有 10 万数据的实验，表明采用中国人名尾字的方法可以得到较好的效果。如果能够收集到更多数据，采用人名中间字和尾字作为整体的方法能够得到更好的效果。

3　使用上下文的人名性别识别

使用上下文特征的人名性别识别需要考虑两个问题：一是如何选取有效的上下文特征，使得这些特征对于人名性别的识别起到很好的效果；二是如何应用已经选择出来的上下文特征进行人名的性别识别。下面分两部分来进行分析。

3.1　上下文特征的选择

日常用语中，“先生”、“小姐”、“女士”、“帅哥”等都能用来作为指示人名性别的上下文特征。我们采用了两种策略来获取性别指示词：从 Hownet 中获取，和利用网络挖掘。

3.1.1.　从 Hownet 获取性别指示词

英文中判断词语的性别属性时可以采用在 WordNet 中检索的方法[11][12]。Hownet 是一个中文信息知识库[13]，描述概念、概念之间的关系和概念所具有属性之间的关系。

Hownet 知识元如图 1，其中“W_C”表示中文词语。DEF 表示定义，其中“male|男”表示该词是男性指示词，如图 1.a；“female|女”表示女性指示词，如图 1.b；“#male|男”表示男性相关指示词，如图 1.c；“#female|女”表示女性相关指示词，如图 1.d。

W_C=爸爸 DEF=human|人,family|家,male|男	W_C=伴娘 DEF=human|人,#GetMarried|结婚,female|女	W_C=男篮 DEF=fact|事情,exercise|锻练,sport|体育,#male|男	W_C=别针 DEF=tool|用具,*decorate|装饰,#female|女

a. 男性指示词（共 199 个） b. 女性指示词（共 397 个） c. 男性相关指示词（共 9 个） d. 女性相关指示词（共 62 个）

图 1 Hownet 中抽取得到的性别指示词示例

Fig.1 Gender designator exapmles extracted from Hownet

3.1.2. 网络挖掘性别指示词

Hownet 的性别指示词是基于人经验编写的，没有包含目前流行的指示性别词，例如“超女”等，而且没有很好的领域扩展性。因此，我们尝试网络挖掘来获取性别指示词。

百度风云榜[2]包含指明男性的帅哥、男歌手、男明星风云榜；指明女性的美女、女歌手、女明星风云榜。各风云榜包含排名最靠前的 50 个人物。抽取 2006 年 4 月 14 日的这六个排行榜，去除重复以及不是中国人名的词条，得到 93 个男性名人、88 个女性名人。随机抽取 100 个人名（50 男、50 女）用于性别指示词语的获取，剩下 81 个（43 男、38 女）用于性别识别实验的测试。

分别将 100 个人名放入百度，将返回 Snippet 分词、词性标注。对 Snippet 的各个性别相关类型词[3]进行统计。随后统计每个词在描述男、女性的 Snippet 中的出现情况。选择词语时，为了综合比例和频次因素，采用贝叶斯参数学习中的 Beta 分布的方法[14]。一个词语，用于男、女性的次数分别为 α-1 和 β-1，对应的 Beta 分布为[α,β]，该分布的方差可以用于描述 α 和 β 的突显性，如公式（2）。对于相同比例的 α 和 β，数据规模越大对应方差越小。例如 Beta[3,2]和 Beta[300,200]对应的方差分别为 0.04 和 0.000479。

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{2}$$

选择性别指示词时，抽取性别比例大于 75%的词语，对它们按突显性排序（从小到大）取出前 20%。根据前面知道性别的 100 个名人人名，得到男、女性别指示词分别为 143 和 151 个，部分结果如表 4 所示。

表 4 从百度抽取得到的部分性别指示词

Tab. 4 Some gender designator extracted from Baidu

性别	词语	总数	单独次数		比例		突显性
			男性	女性	男性	女性	
男	郎	567	558	9	98.41%	1.59%	3.03E-05
	任贤	562	552	10	98.22%	1.78%	3.38E-05
	恒	392	386	6	98.47%	1.53%	4.42E-05
女	桑	493	1	492	0.20%	99.80%	8.11E-06
	圆圆	511	4	507	0.78%	99.22%	1.88E-05
	容	326	3	323	0.92%	99.08%	3.66E-05

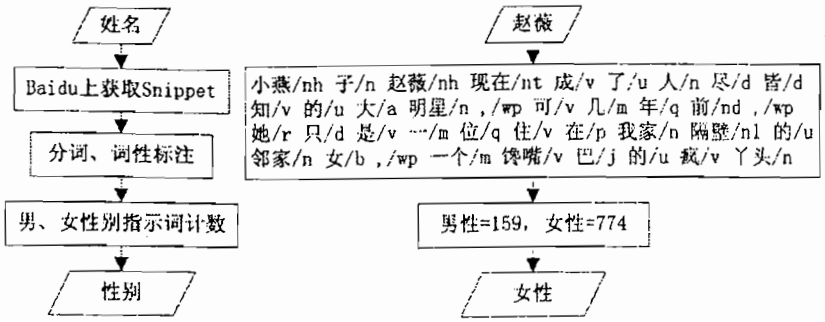
对比前面从 Hownet 中获取的词语，发现 Hownet 中对于一些词语的 DEF 中不包含相应的性别指示信息。例如 Hownet 中“性感”的 DEF 是“DEF=aValue|属性值,appearance|外观”；“女郎”的 DEF 是“DEF=human|人,female|女,adult|成”。

3.2 采用性别指示词的上下文人名性别识别

参考 3.1.2 获取性别指示词的方法，采用性别指示词的上下文性别识别流程如图 2：

[2] http://top.baidu.com/

[3] 这里性别相关类型词包括普通名词、动词、形容词、副词、代词、习用语、机构名、处所名词和地名。



a. 基于Baidu的性别识别流程　　b. 流程举例

图 2　基于 Baidu 的性别识别流程

Fig.2 Gender recognition flow based on Baidu

对 3.1.2 中 81 个测试人名，分别采用 Hownet 和百度抽取的性别指示词实验，结果如表 5 所示。

表 5 基于指示词的上下文人名性别识别结果

Tab. 5 Gender recognition experimental result on designator in context

交叉矩阵	Hownet 指示词		百度指示词	
	Stand-Male	Stand-Female	Stand-Male	Stand-Female
Baidu-Male	41	0	38	0
Baidu-Female	0	37	3	37
Baidu-Unknown	2	1	2	1

表 5 中 Stand 表示标准答案，Male 表示男性，Femal 表示女性，Baidu 表示根据 Baidu 返回的 Snippet 判断的结果。由于是比较两种性别指示词的数量，当出现二者相等时，性别未知(Unknown)，表中数字表示对应人名个数。Hownet 和百度指示词实验中 Stand-Male、Baidu-Unknown 都是 2，对应都是“伍佰”和“谢霆锋”。详细查看，它们分别被标注为“伍/nh 佰/v”和“谢霆/nh 锋/n”。这个错误使相关 Snippet 都被判为无效。两个实验中 Stand-Female、Baidu-Unknown 都是 1。原因都是“孙俪”被标注为“孙/nh 俪/v”了。人名词性的错误在这里产生了一些影响。

两个实验中只有百度指示词实验下有被错判的情况。Stand-Male、Baidu-Female 对应的 3 个人名分别是“张学友”、“赵传” 和“朱智勋”。这是因为前两人有很多爱情歌曲，其中很多词汇关于女性，“朱智勋”是韩国影星，有一些爱情方面的电影。用于抽取特征词的 100 个名人中，对于爱情主题女性名人的上下文更多一些。因此爱情主题大量出现在描写男性的上下文中会误导性别的判断。

基于上下文特征的性别识别结果，整体准确率是性别正确识别的人名个数占总人名个数的比例。表 5 中 Hownet 和百度指示词实验的整体准确率分别为 96.3%、92.6%。

Hownet 性别指示词是从人的经验获得的，而网络挖掘方法完全是基于统计的。两个实验效果差不多，说明网络挖掘性别指示词的方法可以取代基于 Hownet 的方法。百度风云榜上抽取到的人物集中在娱乐界，因此抽取到的词汇多是娱乐领域的。如果加上其他领域的名人应该会得到分布更加均衡的性别指示词。这种方法可以抽取更加适应具体领域的性别指示词。从 Hownet 抽取到的性别指示词在娱乐界人物上可以很好的进行性别识别，说明这些指示词具有较好的通用性。

4　结论和展望

本文尝试了两种中国人名性别的自动识别方法。基于人名用字的实验显示在训练数据只有几万规模的时候，仅采用人名尾字的方法具有最好的实用性。从 Hownet 和互联网挖掘得到的性别指示词，在进行基于百度的性别识别中，都起到了性别指示作用。证明语言对不同性别的描述具有不同的倾向性。

下一步工作，我们将集中在上述两种方法的结合上，即综合利用人名用字和人名上下文来完成单篇文档中的人名性别识别，可以考虑将 Hownet 上的性别指示词和网络挖掘的性别指示词融合在一起，使其能够更好的进行基于上下文的性别识别。

参考文献：

[1] 王厚峰.指代消解的方法和实现技术[J].中文信息学报,2000,16(6):9-17.

[2] 梁茂成，李刚. 英汉机器翻译中人称代词的处理[J]. 中文信息学报 1999 (04).

[3] Hua-Jun Zeng. Qi-Cai He, Zheng Chen, Wei-Ying Ma and Jinwen Ma. Learning to Cluster Web Search Results[A]. SIGIR 2004: 210-17.

[4] Kennedy. Christopher and Branimir Boguarev. Anaphora for Everyone: Pronominal Anaphora Resolution Without a Parser[A]. Proceedings of the 16th International Conference on Computational Linguistics 1996.

[5] Lin. Dekang. Dependency-Based Evaluation of Minipar[A]. Workshop on the Evaluation of Parsing Systems, Granada. Spain 1998.

[6] Bergsma, Shane. Automatic Acquisition of Gender Information for Anaphora Resolution[A]. Canadian Conference on AI. 2005: 342-53.

[7] 钱进. 姓名用字的性别差异统计分析[J]. 常州工学院学报, 2004,17 (5): 60-62.

[8] 钱进. 语言性别差异研究综述[J]. 甘肃社会科学, 2004 (6): 47-50.

[9] 董银秀. 语言中的性别因素[J]. 兰州工业高等专科学校学报, 2004.11 (1).

[10] 王厚峰，梅铮. 鲁棒性的汉语人称代词消解[J]. 软件学报, 2005, 16(5):700-707.

[11] Cardie, Claire and Kiri Wagstaff. Noun Phrase Coreference As Clustering[A]. Joint SIGDAT conference on Empirical Methods in NLP and Very Large Corpora (ACL'99), University of Maryland, USA. 1999.

[12] Siddharthan. Advaith. Resolving Pronouns Robustly: Plumbing the Depths of Shallowness[A]. EACL Workshop on The Computational Treatment of Anaphora, Budapest, Hungary, 14 April 2003.

[13] 董振东，董强. 知网和汉语研究[J]. 当代语言学 2001 (01).

[14] Stuart Russell 著，姜哲，金奕江等译. 人工智能——一种现代方法[M]. 第二版，北京：人民邮电出版社，2004 年.557～558 页