面向大型叙事作品的指人成分识别

钱小飞 陈小荷 董宇 何晓丽

（南京师范大学文学院 南京 210097）

摘 要： 本文定义了指人成分的概念，分析了指人成分的构成和分布特征，并面向大型叙事作品，提出了一种基于邻字熵统计和规则发现相结合的指人成分识别方法。实验对小说《英雄出世》的生文本进行了多次抽样测试，取得了 86.93%的正确率和 91.83%的召回率。

关键词： 指人成分；识别；邻字熵；规则

The Recognition of the Component Referred To Person Facing Narrative Works

Qian Xiaofei, Chen Xiaohe , Dong yu， He Xiaoli

(School of Chinese Language and Literature，NanJing Normal Univ，NanJing 210097)

Abstract: The paper defines the concept of the component referred to person , and analyzes its composing and distributing feature .Based on the analysis , it advances a method to recognize the component referred to person ,which bases on neighbouring character entropy and rules . The experiment sample the raw test of novel Hero Borns to test for times , and it acquires an accuracy of 86.93% , while an recall rate of 91.83%.

key words: component referred to person；recognition；neighbouring character entropy；rule

1 引言

未登录词问题是自动分词的一大难点。在真实文本中，人名在未登录词中占据了较大的比重，成为词法分析关注的焦点之一。同时，人名作为命名实体，它的识别是指代消解的前提，对于提高自动文摘的精确率也有着重要意义。

人名识别大致有两种方法：基于规则的方法和基于统计的方法，在实际应用中常常将二者结合起来。基于规则的方法通过上下文限制性成分（宋柔，1993），姓氏用字分类（孙茂松，1995）等信息进行姓名识别；基于统计的方法（郑家恒，1999）被当前的大多数系统所采纳，它主要通过计算姓氏用字和人名用字的频率，取得人名的概率信息来识别中文姓名。

总的看来，无论是规则方法还是统计方法，以往的人名识别研究表现出以下一些特点：1. 主要是针对于姓名的识别，而非整体意义上的指人命名实体的识别；2. 充分利用了疑似串作为姓名的内部结构信息以及上下文信息；3. 常常采用新闻语料作为测试语料。

本文在前人研究的基础上，定义了指人成分的概念，并从篇章统计的角度，面向大型叙事文本设计了一种基

作者简介：钱小飞 （1981-），男，江苏南通人，硕士生。E-mail：qierflying@163.com .

于邻字熵统计和规则方法相结合的指人成分的识别方法，并分析了这种方法的适用性。

2 指人成分的构成与分布

2.1 指人成分的构成

目前中文人名的识别主要局限于姓名的识别。中文姓名的构成形式可以分为完全形式和非完全形式。完全形式是指“姓（1~2 字）+名（1~2 字）”的形式。其中姓氏有单姓、复姓和双姓之分。人名有单名和双名之分。非完全形式包括有姓无名、有名无姓、前缀+姓、姓+准后缀、姓+身份词等形式。

姓名是指人成分的主体。对姓名进行结构分析有利于构造有效的识别算法，如当前广泛采用的通过姓氏用字和人名用字的概率计算姓名概率的方法。但另一方面，构造一个健壮的人名识别系统，加强对不完整形式的人名和其他形式的指人成分的识别也十分有必要。这些边缘问题常常更难以发现特征，并在一定条件下可以转变为主要问题，如在小说《英雄出世》中，由于作为小说主要人物，“霞姑”[1]，这一没有姓氏的人名形式成为一个高频人名。

我们把指人成分（CP）定义为在篇章范围内指人的命名实体，满足指称性、专门性、词汇性、开放性等命名实体特征。其中指别性（指称性和专门性）是 CP 的根本特征。词汇性则规定了它区别于短语的特征，如它不包括含指代词的短语，“他哥哥”，“我的叔叔”之类不属于 CP 范畴。

具体说来，可以将中文指人成分分为以下三类：

第一类是中文姓名，包括了上文提到的完整形式和不完整形式。如“张华”、“小王”、“赵总”、“陈叔叔”。

第二类是人物的别名和绰号，在小说文本中有高频出现，作为对人物某种特征的介绍和隐喻，常常用事物来命名，如“神雕大侠”，“及时雨”，“小玉兰”，“店小二”。

第三类是其他在整个篇章范围中具有指称性和专门性的指人形式，如按排行的称谓“老五”，“十四爷”。

其中，第二类和第一类中有名无姓的人名形式具有较大的开放性，是 CP 识别的难点。第一类中“姓+身份词”、“单名+身份词”的形式，身份词随着文本的内容的变化，可以体现出不同的时代特征和领域特征，很难搜集完全，如“管带”，“镇守使”现在已不再使用，完全依靠规则的方法也难以胜任识别任务。

表 1 以“姓+名”形式作为区分，统计了《英雄出世》中 CP 的构成状况：

表 1 CP 构成表

字段 / CP 形式	用例数（例）	比例（%）	词型数（个）	比例（%）
“姓+名”形式	2623	26.30	33	19.41
其它形式	7349	73.70	137	80.59

从 CP 的形式和所指的关系来说，一个 CP 所指可以对应多个指人形式，如《英雄出世》中，巴庆达对应着“巴庆达”，“巴哥哥”，“小巴子”，“老巴”，“巴”四种形式。这些形式也可以表现为无用字联系的形式，如某个人物的一个或多个别名。

可见，CP 的概念包含了中文姓名、人物别名等多种指人形式，同时，它也打破了姓名用字和长度（2~4 字）的规律，增加了识别的难度。

2.2 指人成分的分布

经典的人名识别方法常常使用人名的上下文信息，宋柔（1993）提出了使用限制性成分识别人名的方法，适用这一方法的依据是“在日常见到的语料，尤其是新闻语料中，首次提到一个名不见经传的人名时，一般要在人名前或后加一些限制性成分，作为作者对这个陌生人认知的出发点。以后再提到此人时，便可不加限制性成分而只提其名。”

这一认识主要是基于新闻语料的观察。离开新闻语料，比如在口语语料，文学语料中，这种说法是否还适用

[1] 霞姑在小说中是人名而非“霞+姑”的“人名+准后缀”的构词形式。

值得商榷；并且，作为身份词是否是已登录词也未可知。我们据《英雄出世》进行了小规模统计，小说第一章中共出现 CP 型数 19 个，其中前后出现身份词的 CP 型数 8 个，占 42.11%，其中包含 3 个“姓+未登录身份词”的 CP 词型。

与新闻语料相比，大型叙事作品中指人成分在分布上也表现出一些新的特点。

首先，对人物的介绍（限定词或短语）较少，因此 CP 分布位置相对趋于句首。

其次，人名的后邻接词的不确定减小，表人物动作的动词，如“说”、“道”、“笑”，以及一些心理动词，如“觉得”，“认为”等，常常在后邻接位置上出现。这是由于在叙事性作品中，这些动作是构成人物交流的基本动作。

最后，大型叙事作品是情节驱动的篇幅较长的作品，围绕此情节出现主要人物和次要人物，因此单个人名有较高的出现频次。

我们分别以 CP 分布的句首概率、后邻接动词概率[2]、平均频次三个统计指标描述以上三个特点，对小说《英雄出世》进行了统计，表 2 给出了相关统计数据：

表 2 CP 分布特征统计表

CP 总频次	9972	CP 总频次	9972	CP 总频次	9972
句首频次	3370	后邻接动词频次	6705	CP 词条数	170
句首概率	33.79%	后邻接动词概率	67.24%	平均频次	58.66

从中可以看出，CP 在频次和出现位置上都表现出比较明显的特征，高频特征有利于使用统计的方法进行识别，位置特征有利于制定针对性的规则加以发现。

3 统计与规则相结合的 CP 识别

3.1 识别任务

根据上文提出的 CP 的概念，指人成分识别的目标是在线性文本中标注出中文指人命名实体，包括中文姓名，人物别名等形式。根据这一概念的构成，可以将识别任务分解为姓名识别和其他人名的识别。

本研究分别用“【”和“】”标注中文人名的左边界和右边界。以周梅森的小说《英雄出世》生语料作为测试语料，标注结果示例如下：

昔日【百顺】、【玉环】、【老五】和【方营长】一起来过的。

3.2 识别方法

根据指人成分的结构和分布特点，我们将基于邻字熵的统计方法和基于邻接词的规则方法结合起来进行 CP 识别。

3.2.1. 基于邻字熵的识别方法

熵是对单个随机变量不确定性的度量。字符串的邻字熵描述了该串的上文或下文的自由度。基于邻字熵的方法将人名看成是上下文自由，内部结合紧密的语言单位。邻字熵的计算公式如下：

$$H(LZ) = -\sum_{zi \in ziSet} p(zi)\log_2 p(zi) \qquad (1)$$

其中，p(zi)是随机离散变量 LZ 的概率密度函数，zi 属于字符集 ziSet.

通常如果一个字符串的左右邻字熵较高，而其非临界位置的子串左右邻字熵较低，可以将该字符串识别为未登录词。基于邻字熵的未登录词识别可以描述为触发和维系两个过程，分别用触发邻字熵和维系邻字熵来表示。

触发是 CP 识别的开始，触发邻字熵是指触发 CP 识别的字符串左邻字熵和右邻字熵组合，包含触发左邻字熵

[2] 此概率通过 CP 词例后单字串和双字串与动词词表比对计算得出。

和触发右邻字熵两个部分。

维系是寻找 CP 右边界的过程，在满足触发条件的基础上，维系邻字熵是用来寻找 CP 右边界的字符串左右邻字熵组合。

该识别方法就是在触发和维系两个动作的基础上构建起来的，具体步骤如下：

第一步，对待识别的文本建立 n 张串邻字熵表，第 i 张表存储了长度为 i 的字符串及其出现频率、左右邻字熵等信息。其中 $1 \leqslant n \leqslant 3, 1 \leqslant i \leqslant n$. （A）

第二步，顺序扫描待识别文本，对文本中的每一个位置，依次获取长度为 len（len≥2）的字符串 str. （B）

第三步，循环进行 i 元触发-维系识别：（C）

- 长度检查，如果字符串长度 len≤i，转 B.(D)
- 触发，如果字符串的左邻字熵 HL(str)大于左触发邻字熵 HLT[i]，并且字符串右邻字熵 HR(str)小于右触发邻字熵 HRT[i]，转 F；否则，i++,转 C.(E)
- 特征-频率检查，如果字符串起始字符 str[0]∈姓氏用字集合 nameZi 并且字符串频率 freq(str)>低频阈值 Freq1 [i]；或者字符串频率 freq(str)>高频阈值 Freq2 [i]，转 G；否则，i++,转 C. （F）
- 维系，如果字符串的左邻字熵 HL(str)大于左维系邻字熵 HL1[i]，转 C；如果字符串右邻字熵 HR(str)大于右维系邻字熵 HRI[i]，将从识别起始位置开始至 str 结束的字符串写入人名表，否则 str 右移一个字符，转 G. （G）

第四步，分析 CP 表，过滤高频常用词，建立姓氏表和名字表。(H)

第五步，利用 CP 表，姓氏表以及名字表进行基于最大匹配的 CP 标注。（I）

一般而言，在真实文本中，字符串的长度与其出现频率和邻居种数构成反比关系，因此在 i 元触发-维系识别中，i 值越大，str[i]的邻字熵均值越小。从统计本身来说，str[i]的邻字熵均值越小，识别的可信度越高；但统计方法对于低频事件的估值常常欠准确，n 的取值也并非越大越好，所以我们限定了 n 的取值范围为[1，3].

从语言学的角度考虑 i 元触发-维系识别的问题，由于字和词是汉语的一些基本单位，字是汉语最自然的分隔单位，词是最小的能够独立运用的语言单位，从而在组合关系中，它们也必然有较大的外接触面（邻居种数多），而汉语绝大部分词是双字词，所以基于一元、二元和三元的触发-维系识别的发现能力是递增的。因此，在实际 CP 识别过程中，二字 CP 名和三字 CP 分别基于一元和二元触发-维系识别来发现，四字及更长的 CP 基于三元触发-维系进行识别。

3.2.2. 基于规则的识别方法

基于规则的方法有两个目标，一是确认基于邻字熵方法的识别成果，提高识别的正确率，二是发现单字人名及其他一些基于邻字熵方法未能发现的 CP，主要是低频 CP。

如上文分析，大型叙事作品常常在句首用一些简单的言语形式来表现人物的一些基本动作和心理感受，如“霞姑道”，“边义夫说”等。此外，一些表示人物身份、地位的词也会出现在姓氏、人名的前后，参与部分人名的指别和称谓，如“李太夫人”，“钱管带”等。为此，我们建立了动作词表、副词表和身份词表，参与规则方法的识别。

动作词表和副词表从《英雄出世》人名标注语料中获得。本文选取该小说第 26，32，38，44，45，50， 55，58 章（这些章节与下文的测试语料不重合），通过比对词表的方法，自动提取出比邻于句首位置的人名后的动词或副词，并经过人工甄别和简单拓展，共获得 125 个表现人物基本动作和感受的动词和 48 个出现在该位置的副词，分别加入动作词表和副词表。

身份词表通过内省的方式获取。我们从亲属、学校、军队、政府官员等系统的角度进行了常用身份词的简单归纳，获得了 58 个身份词。《英雄出世》中的身份词大致覆盖了其中亲属和军队方面。其中军队系统中有许多时代色彩较强的词汇，如“标统”，“天帅”等，身份词表不加以收录。

规则的形式有两种：第一种形如“<0>XCP/[Adv]/V”，表示当疑似 CP 串 XCP 位于句首，其后出现指定动词 V 或副词 Adv 和动词 V 时，将 XCP 识别为 CP；第二种形如“XCP/S”，表示当 XCP 后紧邻身份词 S 时，将 XCP 或 XCP+S 识别为 CP. 这两条规则通过动作词表、副词表和身份词表来求解 CP. 识别步骤如下：

第一步，顺序扫描待识别文本句子，对句子的每一个位置，进行如下识别过程：（J）

- 查姓氏表，如果查找成功，对其邻接单元查身份词表，如果查到，将姓氏与身份词的邻接形式加入 CP 表；（K）
- 如果该位置位于句首，在其 4 字右邻接单元中，查动作词表，如果查找成功，查副词表，过滤动作词前字符串的尾部副词性单位，如“又”，“却又”等，将该字符串加入 CP 表；（L）
- 否则，对句首位置依次查名字表和姓氏表，如果查找成功，对其邻接单元查身份词表，如果查到，将单名+身份词、姓氏+身份词的邻接形式加入 CP 表。（M）

第二步，分析 CP 表，过滤高频常用词，建立姓氏表和名字表。（N）

第三步，利用 CP 表，姓氏表以及名字表进行基于最大匹配的 CP 标注。（O）

基于规则的方法利用了基于邻字熵的方法的一些识别成果，如 CP 表，实验中我们将两者结合起来（I 步骤和 O 步骤归一），进行统一的 CP 标注。

4 实验结果

实验以周梅森的《英雄出世》生语料作为测试语料。为验证算法的识别效果，将其分为 33 个样本，每相邻两章为 1 个样本（第 47 章为小说第二部分末章节，归入样本 23），进行多次抽样测试。

我们通过实验的方法对包括触发邻字熵、高频阈值、低频阈值进行了估值，随机抽取 5 个样本，统一设置阈值，当满足条件

$$\begin{pmatrix} HLT[i]>0.1,\ i\in\{1,\ 2,\ 3\} \\ 0.0\leqslant HRT[i]\leqslant 1.2,\ i=1 \\ 0.0\leqslant HRT[i]\leqslant 0.5,\ i\in\{2,\ 3\} \end{pmatrix} \&\& \begin{pmatrix} HRI[i]>0.1,\ i\in\{1,\ 2,\ 3\} \\ 0.0\leqslant HLI[i]\leqslant 1.4,\ i=2 \\ 0.0\leqslant HLI[i]\leqslant 1.0,\ i\in\{1,\ 3\} \end{pmatrix} \&\& \begin{pmatrix} Freq1[i]>1,\ i\in\{1,\ 2\} \\ Freq1[i]>3,\ i=3 \\ Freq2[i]>3,\ i\in\{1,\ 2\} \\ Freq2[i]>4,\ i=3 \end{pmatrix}$$

时，实验结果如表 3 所示：

表 3 《英雄出世》抽样实验结果

字段 / 抽样	所属章节（章）	识别数（例）	正确数（例）	实有 CP 词数（例）	正确率（%）	召回率（%）
样本 1	1~2	208	183	219	87.98	83.56
样本 5	9~10	285	245	257	85.96	95.33
样本 18	35~36	369	322	335	87.26	96.12
样本 21	41~42	262	222	244	84.73	90.98
样本 33	66~67	231	205	220	88.74	93.18
均值评价	####	271.00	235.40	255.00	86.93	91.83

从识别结果看，基于邻字熵的方法和基于规则的方法对于人名的识别相互确认，互为补充。前者对于相对高频的人名有着较好的发现能力，后者对于低频，或长度小的人名发现效果较好。如，样本 1 中，CP 形式“边哥”通过查姓氏表和身份词表的方式发现得以识别，这是基于邻字熵的方法未能发现的。

实验中所出现的错误主要来自三个方面。一是由于统计方法处理低频事件时发生的错误和遗漏，如样本 5 中低频串“毛瑟”识别为 CP；二是一些单字指人成分同时也可以作为单字词存在，实验中没有将它加入 CP 表，导致了单字 CP 的识别失败，如样本 1 中大量出现了“边义夫”的姓氏“边”构成的单字 CP，造成召回率低下；三是一些多次出现的未登录词容易被误识，如样本 18 中，“独香亭茶楼”出现 5 次，进入 CP 表，导致识别错误。

5 结语

本文分析了大型叙事文本中指人成分（CP）的构成与分布特征，基于这些特征设计了一个基于邻字熵和规则相结合的人名识别方法，并进行了初步实验。与经典的人名识别方法相比，这一方法没有从分析人名用字的角度

的构造算法，而侧重于从文本统计的角度识别 CP，较好地识别了非常规指人单位，以及其他通篇缺少姓氏的人名类型。对于常规的“姓+名”形式，我们仍然主张使用经典的通过用字频率计算人名概率的方法，通常它更好地化解了低频事件的问题。下一步我们考虑将该方法与经典的人名识别方法结合起来，并基于大规模语料全面地归纳和统计 CP 的构成规律、分布规律以及用字规律，加强规则的描写，以提高识别精度。

参考文献：

[1] 刘秉伟. 基于统计方法的中文姓名识别，中文信息学报，第 14 卷，第 3 期。

[2] 罗志勇. 一种基于可信度的人名识别方法[J]. 中文信息学报，第 19 卷，第 3 期。

[3] 孙茂松. 中文姓名的自动辩识[J]. 中文信息学报， 第 9 卷，第 2 期。

[4] 宋柔. 基于语料库和规则库的人名识别法[A]. 陈力为，袁琦主编. 计算语言学进展与应用[C]. 北京：北京语言学院出版社。

[5] 张华平. 基于角色标注的中国人名自动识别研究[J]. 计算机学报 2004 年，1 月。

[6] 郑家恒. 基于语料库的中文姓名识别方法研究[J]. 中文信息学报，第 14 卷，第 1 期。

[7] 张仰森. 基于姓氏驱动的中国姓名自动识别方法[J]. 计算机工程与应用，2003 年，第 4 期。