基于语料库的数量名短语识别

方芳，李斌

（南京师范大学，南京 210097）

摘 要： 数量名短语的自动识别对用统计方法处理汉语将起到很重要的作用。本文主要是对顺序的“数·量·名”、量词重叠的数量名短语等常见的数量名短语形式自动识别方法的研究，提出了基于 n 个后字匹配的后退算法来发现更多的量名搭配，把召回率提高了 40 个百分点。我们在 240 万字的当代新闻小说语料上进行了识别试验和测试，结果显示，达到 80%左右的调和平均值。

关键词： 数量名短语；自动识别；后退算法；中文信息处理

Copus Based Investigation on MQN Phrase

FANG Fang, LI Bin

(Nanjing Normal University, Nanjing 210097)

Abstract: This paper introduced a new way to automatically identify the MQN phrases based on linguistic rules and statistical method. A corpus about 2.4 million words of news and novels is used for the automatic identification test. To solve the insufficient amount of items in coordinate dictionary, we used back-off algorithms to get more items of QN coordinates based on n-last matching, which raised the recall rate by 40%. The automatic identification achieves about 80% in F-score, which is close to the practical level.

Keywords: Chunk of “Number- Quantifier -Noun”; Automatic identification; Back-off algorithms; Chinese Information Processing

1 引言

自动识别研究最大的难题基本集中于两个方面：未登录词和搭配冲突。我们试图不依靠复杂的句法分析，而是通过构建针对特定任务的专门词典和概率统计模型克服未登录词和数据稀疏，从而解决数量名短语识别及量名搭配决策问题。基于相同末字的后退算法是本文的主要创新之处，通过从搭配词典中自动获取相同末字构建名词后退词典，自动获取量词构建量词后退词典（即下文的量词类型词典），在识别过程最多可达到三步后退的办法，来召回搭配词典中未登录的量名搭配。此外，在加快算法速度方面，我们还采取了建立双外索引的办法，优化软件效率。本文是对数量名短语识别算法的初步探讨，实验证明还有一定的研究空间。

2 研究内容

我们的研究对象是这样的数量名短语形式：

指示代词/数词/（指示代词+数词）+量词成分（QE）+名词中心语（NE）

基金资助：南京师范大学 211 资助项目 语言信息处理与分领域语言研究的现代化（1240702504）

作者简介：方芳(1981-)，女，安徽巢湖人，硕士。E-mail：fredagrape@Gmail.com.

其形式化描述为：

MQN -> (P+)NUM|P+QE+NP

NP -> 名词|名词短语

◆ P表示指示代词，即这、那。NUM表示数词，包括各类数词、连接多个数词的词语和表概数的辅助词语，如到、至、来等；

◆ 如果前一个P已出现，则后一个P必不出现；

◆ QE包括单个量词和量词重叠式；

◆ NP指单个名词或各种形式的名词短语，也包括数量名短语；

◆ 当量词后仅有一个词语且与量词形成搭配的话，NE即指该词语。当量词后有一个名词短语且其名词中心语与量词形成合理搭配的话，NE指该名词中心语；

3 研究重点、难点及对策

MQN 的识别重点和难点与一般的名词短语识别不同。后者的定语部分情况复杂，边界形式标记又不充分，因此左边界特别不容易判定。而在 MQN 中，数量结构是中心名词最外一层的修饰语，且一般处于中心词的左边；加之量词基本是一个封闭集合，数词的表达方式又明显有别于汉语其他词类，这就使得左边界难以确定的问题并不凸显。相反地，量名搭配的多样化和复杂性给 MQN 右边界的确定带来重重困难，尤其是在数量结构与终止符[1]之间出现多个名词的时候。量名搭配的确定毫无疑问是识别的重点和难点。因此，MQN 识别的关键就在于是否能有效地对量名搭配做出准确选择和判定。

3.1 数词分析

我们从形式和组合规律上对汉语数词作出以下分类：

✧ 基数词——阿拉伯数字（串）的汉语表达方式，包括小数、分数。

例：一、二十九、四点五三、十分之一、三百五十六、百、一百

✧ 序数词——“第”与某些基数词（除却小数、分数和百、千、万、亿的基数词，下同）的顺序组合。“初”虽然也能与某些基数词顺序组合，但是一般不出现在量词前面，因此不予考虑。

例：第一、第五十三

✧ 特殊数词——以非基数词的汉字表示数量多少、程度、范围的形式。

例：数、头、首、整、双、半、几、多、无数、好几、若干、大半、多少、成千上万、上千万、上百万

其中，大部分特殊数词直接与量词顺序连接使用，如：数（名）（官员）、整（个）（城市）。但半、多可以跨越量词而与某些基数词组合使用，如：一(个)半(科学家)、三(个)多(星期)。多还可以跨越量词而与半组合使用，如：半(个)多(世纪)。

✧ 基数词（不含小数、分数）和特殊数词的组合

例：数十(位)、头几(场)、头三(个)、五百多(页)

一般而言，能与基数词组合的特殊数词只有数、头、多。

✧ 基数词（不含小数、分数）与余、来的组合

例：三十余(所)、十来(个)

3.2 数量结构分析

数量结构是汉语中主要的物量表达方式，本文中，我们只讨论那些位于所修饰的名词语前面的数量结构，主要有以下七种结构形式：

MQ1：单个数词+单个量词（+多/半）

例：一条街、5篇论文、首场大雪、两项省部级以上的科研项目

MQ2：单个数词+量词重叠式

例：一个个平凡而真实的日子、一幅幅艺术作品

MQ3：数词复叠式+单个量词

[1] 终止符——指一个句子或小句的结尾符号，包括，。！？；等。

例：3到6个月、第四、五、六届全国委员会。

MQ4：数量结构重叠式

例：一片一片、一幅又一幅

MQ5：数量结构复叠式

例：5级一6级偏北风、几十个甚至上百个国家

MQ6：量词重叠式

例：座座/m　青峰/n

MQ7：单个数词+形容词+单个量词

例：一大摞贺卡、一大麻袋还带着泥土芳香的花生

与数量结构相比，指量结构的形式则简单得多。主要有两种结构形式：

PQ1：指示代词+QE

例：这项政策

PQ2：指示代词+MQ

例：这两个项目

数量结构和指量结构在语料中的分布情况如下表。

表 1　　数量/指量结构在 98 年 1 月语料中的分布比例

单位：条	数量名短语	数量结构	指量结构
各类型条数	13091	10215	2876
占总数比例	100%	78%	22%

3.3　量名冲突

由于名词或名词短语间的冲突，使得每条量名搭配在语言学上和算法上呈现出不同的状态，主要表现为语言学上的歧义关系和算法中的竞争关系。

从语言学角度看，造成了量名搭配的三种歧义形式（IPR）：

IPR1：数量/指量结构之后、终止符之前有多个名词，而实际上只有一个能与量词形成合理搭配。

例：给[这项/r 活动/n]筹措更多的 资金/n

该句中的量词“项”与之后的名词“活动”能够形成一条合理的量名搭配，而与名词“资金”则不能够。判断的标准来自“项”这个量词本身对名词的选择，以及“活动”和“资金”对修饰自身的量词的选择。对机器来说，如果没有先验的知识，所有的名词没有任何区别。我们专门构建了一部的量名搭配词典中，让计算机通过查对搭配词典基本消解此类歧义。

IPR2：数量/指量结构之后、终止符之前有多个名词，且它们有的或全部都能与 QE 形成合理搭配。但只有一个名词与量词在同一个句法层面上。

例：[多种/m 类型/n]的 帮困/vn 服务/vn 小组/n

该句中的量词“种”是一个泛指量词，一般情况下绝大多数的名词都能与其形成合理搭配，然而，对这个短语进行结构分析，我们可以看到，“多种类型”这个数量名短语是“帮困服务小组”的修饰语。这几个看似合理的量名搭配中，只有“类型”才是这条量名搭配的中心词。这种歧义的消解需要在句法层面判定。目前普遍使用的句法分析器一般难以克服这个问题，因此我们试图引入边界概率，来解决此类问题。

IPR3：选择哪个名词与量词搭配，有时即使在句法层面也无法判定。（由于基本不影响理解和应用，如机器翻译等，因此不予进行歧义消解）当然，选择的结果对语意侧重及感知有一定影响。

例：[一些 登山者/n]的 足迹/n

[一些 登山者/n 的 足迹/n]

若拿计算语言学的术语界定打比方，这第三种歧义可以称作“伪歧义”。因为，该句中的量词与其后的两个

名词都能形成合理的搭配，并且无论是哪种搭配，似乎都不影响整句语意的正常表达，也基本不会造成句法歧义。除非将其放到足够大的语境中去观察语义联系或指向，否则人也很难判断出哪个名词和量词搭配才算是正确的。

从算法的角度看，造成了量名搭配中，名词或名词短语间两种不同的竞争关系，以[第一/m　个/q　有关/vn　军事/n　安全/an　磋商/vn　机制/n　的/u　协定/n]为例：

一是在搭配词典内的竞争：即量词与多个候选名词的搭配在搭配词典内都能找到时，哪一个名词为符合语义语法的正确搭配。如例句中的名词机制和协定。这种竞争关系可对应于 IPR2 和 IPR3。

二是搭配词典外的竞争：即参与竞争的名词有的可在量名搭配词典内找到与该句中量词的搭配条目，有的找不到，如例句中的安全和协定。这种竞争关系可对应于 IPR1。

4　算法描述

我们制定了基本的识别策略。即：先定位 MQN 的左边界；再利用特殊语言规则约束和一般词例知识来排除非数量名短语结构；最后利用量名之间的搭配概率和中心名词的边界概率来进行 MQN 的自动识别。

本文对数量名短语的识别工作分为三个模块进行。

4.1　数量/指量结构归并模块

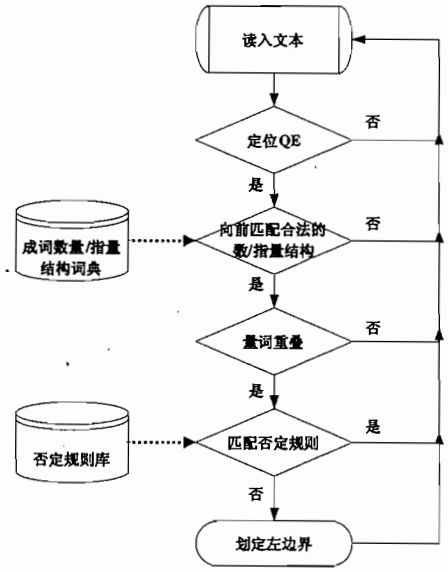

图 1　左边界识别流程

4.2　量名搭配识别模块

该阶段，我们采取了三种策略结合。首先，搭配概率最高的作为中心名词；边界概率次之；肯定规则为判定 MQN 的最后一道标准。在选择的过程中，只有当本条标准没有发挥作用时，才能采用下一条标准。在搭配概率的计算中，我们运用了“后退算法”，即先将名词后退到双字词尾，不成功再后退到单字词尾，还不成功再后退到泛指量词，最后将量词和名词同时后退。之所以依次后退而不是加权于各参数之上，是由量名搭配在语料中的分布规模和特点所决定的，也是对克服数据稀疏的一个尝试。

搭配概率和边界概率的计算模型如下：

我们用 CP 表示搭配概率，用 BP 表示名词的边界概率，Q 表示量词，N 表示名词，i 表示候选量名搭配中的名词在量词到终止符之间所有名词中的第几个。则对每一条候选词串，有如下公式：

$$CP(N_i)=\lambda_1 f(Q,N_i)+\lambda_2 f(Q,N_{i级})+\lambda_3 f(Q_{泛指},N_i)+\lambda_4 f(Q_{泛指},N_{i级}) \quad (1)$$

$$BP(N_i)=\frac{f_{边界}(N_i)}{f_{内部}(N_i)} \quad (2)$$

注：公式（1）是个四项式，依次表示一次原始查询和四步后退查询（见 **3.1.1.2.**）时的搭配概率计算。其中，每项前的 λ 为

加权值。当前后退项的 λ 值为 1，同时其他项的 λ 值为 0。算法优化时，还可考虑对不同项的 λ 值作不同权重处理。

由此，要在候选词串中确定一条量名搭配，我们只需求 $\arg\max_i CP(N_i)$，即可得 i 个候选名词中搭配概率最大者；求 $\arg\max_i BP(N_i)$，即可得 i 个名词中边界概率最大者。

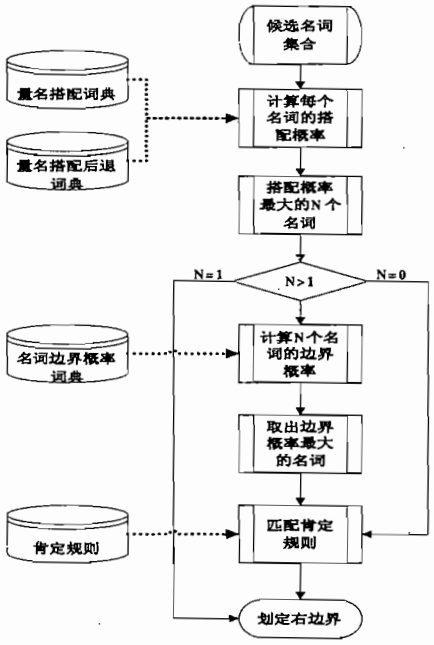

图 2 右边界识别流程

4.3 肯定否定规则

肯定规则和否定规则也被引入，从句法结构上确定或排除不合法的候选结构。

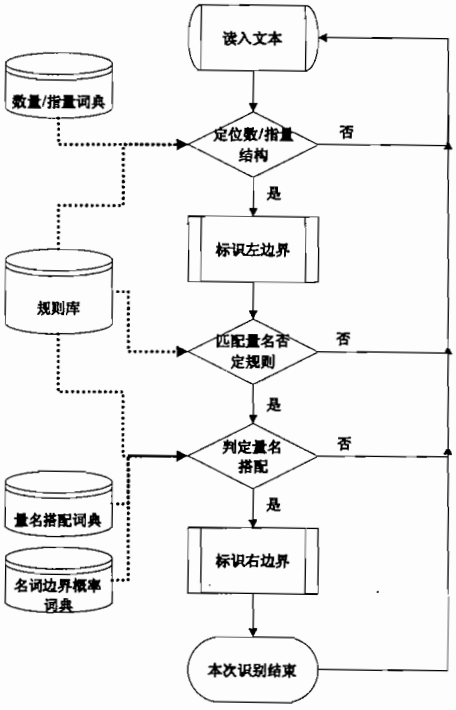

图 3 系统总流程

5 测试

我们的测试分为封闭测试、开放测试和扩展开放测试。

词典资源：我们从1998年1月前20天的《人民日报》语料（Corpus1，规模共691,336词次）中抽取了2108条量名搭配，并统计频率，作为最初的搭配词典（QNDictB）；将两部纸质量名搭配词典（郭先珍，2002；何杰，2001）和一个量名搭配集合[2]作为补充，合并建立一般的封闭和开放测试的搭配词典资源（QNDict）。QNDict中共有搭配13,618条。另外，我们将从98年全年中自动提取的量名搭配和另一部规模较大的纸质搭配词典（刘学敏，1989）与QNDict一起并为QNDictG，用作扩展的开放测试资源，QNDictG共有量名搭配41,556条。依据语言学研究成果和量名搭配中提取的数据，建立量词类型词典（QType），其中收集量词640个，类型4种。边界概率表的数据从训练语料中统计获得。

测试语料：除Corpus1外，还有1998年1月后10天的《人民日报》语料（Corpus2），共约35万词次；以及现代文学家方方的小说语料（Corpus3，经过中科院计算所分词软件的分词和词性标注处理），约32万字。

测试结果的基线（Baseline）是：只利用训练语料获取搭配资源，不利用其他任何数据信息和算法，进行识别的精确率(Acc.)、召回率(Cov.)及调和平均值(F)。

5.1 实验结果

评测项目为目前普遍采用的测试评估标准：精确率、召回率和调和平均值。

表2　　新闻语料测试结果报告

单位：%	Baseline QNDictB	添加后退算法	QNDict	添加规则	添加后退算法	添加搭配概率	添加边界概率
精确率(Acc.)	78.4	75.6	78.4	87.7	78.2	80	80.3
召回率(Cov.)	33.9	70.7	33.9	38.7	76.4	78.1	78.7
调和平均值(F)	47.3	73.1	47.3	53.7	77.3	79	79.5

对于上表，有以下说明：第一次添加后退算法时，不同时利用规则、搭配概率和边界概率；使用QNDict时，只从中获取搭配资源，不利用其他任何数据信息和算法；后四列都是在以QNDict为词典资源的基础上，依次添加。

从表2中可以发现，单纯依靠搭配词典而不使用其他信息，识别出的数量名短语少，但是较为准确；使用后退算法以后，召回率提高了近40个百分点，说明其在对未登录词的识别上，发挥了较大作用，而且精确率几乎没有受到影响；句法规则可以排除掉大量的错误而迅速提高精确率，并且，由于肯定规则的采用，也能回收一些可能被淘汰掉的数量名短语；搭配概率和边界概率一定程度地弥补了后退算法在识别精确度上的不足，通过解决搭配冲突，促成了精确率的提高，并带动了调和平均值的上升。

但是我们也看到，即使是引入了统计数据，识别结果也不能让人十分满意，究其原因，主要来自两方面：后退词典的有效性不够，以及搭配概率和边界概率还存在数据稀疏的问题。而造成这两个方面原因的直接因素，就是量名搭配词典的规模不够大。而数量名短语的识别显然不能够仅靠无限度地扩大词典资源的规模，这就要求我们在算法上力求优化。表3展示在其他特色算法都具备的前提下，分别用**QNDictB**和**QNDictG**进行封闭测试和小说语料的测试效果。

表3　　总体测试结果报告（单位：%）

语料	封闭测试（Corpus1）		开放测试（Corpus2）		开放测试（Corpus3）	
词典	QNDictB	QNDictG	QNDictB	QNDictG	QNDictB	QNDictG
Acc.	70.3	67.6	79.2	78.1	72.2	80.3
Cov.	70.7	72.8	72	78.5	61	83.7
F值	70.5	70.1	75.4	78.3	66.6	82

对长短距离的数量名短语的识别情况报告：测试语料中有“数量结构+名词”的数量名短语3878条，占总数

[2] 该搭配集合由2003级硕士华滢同学友情提供，在此表示感谢。

的 55.8%，自动识别出 3705 条；“数量结构+n 个词（n>=3）+名词”的数量名短语 1921 条，占总数的 27.7%，自动识别出 412 条。由此可见，自动识别的困难主要集中于长距离的数量名短语上。

5.2 错误分析

通过考察系统对标准标注文本和机器标注文本的自动比对结果，我们发现识别错误主要在于以下几方面：

1、仍有部分量名搭配在搭配词典中未登录，这是影响召回率的主要原因。一些未登录名词存在于语料中，导致含有该名词的量名搭配也未登录。如：一/m 节/q 旋律/n、一/m 副/q 以不变应万变/l 的/u 王者/n 风仪/n，等等。在测试语料中，未登录的量名搭配共有 726 条，占 MQN 总数的 21%。

2、搭配冲突解决不够，仍有中心词错判的情况。这种错误占据了 MQN 识错情况的绝大部分，表现为：一是右边界错识的实例大多是候选词串为定中式的 MQN。如：[那个/r 梦/n] 中/f 的/u 她/r 应为[那个/r 梦/n 中/f 的/u 她/r]。这种错误共计大约占 MQN 总数的 10.6%。其二，一些 MQN 的右边界和终止符之间有其他名词。如：[2 9/m 个/q 主要/b 新兴/b 市场/n 的/u 私人/n]应为[2 9/m 个/q 主要/b 新兴/b 市场/n] 的/u 私人/n，这种错误约占 MQN 总数的 9%。

3、语料本身的标注错误造成了自动识别的连带错误。标注错误一般分为分词标注错误和词性标注错误，后者是造成连带错误的主要原因。如：一/m 副/b 嘴脸/n 中的炮应是量词（q），而被误标成区别词（b），机器不按数量结构处理，就会错过这条 MQN。反之亦然。语料标注错误造成的 MQN 识别错误约占 MQN 总数的 2.1%。

4、候选词串较长时，可能把非数量名短语误识为数量名短语。在某个名词已经出现过的情况下，数量/指量结构可以直接指代以该名词为中心语的 MQN 作为句子的主语。如：批准/v 在/p 浦东/ns 开设/v 分行/n 的/u 外资/n 银行/n 已/d 有/v 2 2/m 家/q ，/w 其中/r 9/m 家/q 已/d 获准/v 经营/v 人民币/n 业务/n 中的“9 家”就指代了 9 家公司。而机器仍然把主语（数量/指量结构）之后的部分作为候选词串来处理，这样就容易错误地多识别出了一些 MQN。如上例就被误识成了[9/m 家/q 已/d 获准/v 经营/v 人民币/n 业务/n]。这部分错误约占 MQN 总数的 19.6%。

6 结论和未来工作

本文采用了基于相同末字的后退算法来获取那些不在搭配词典中的量名搭配，结果证明，这种方法是有效的。启用后退算法后，系统比仅仅依靠搭配词典多召回近一倍的量名搭配，也直接使得数量名短语的召回率大大提高。由于后退词典直接在原有的搭配词典上生成，实际上是一种分类的思想，因此这种算法亦可应用于中文信息处理的其他类似工作之中，如其他类型短语的自动识别、特定句式的消歧，等等。

在今后的工作中，我们将针对以上错误改进和优化算法，以使识别系统取得更好的效果。比如：通过构建统计模型来克服量名搭配的未登录问题；丰富规则库，考虑更多上下文的因素；用机器学习的错误驱动方法为结合点，把统计和规则协调共用起来，以取得一个相对最优的效果。

参考文献：

[1] 朱德熙. 语法讲义[M]，北京：商务印书馆，1982.

[2] 李宇明. 汉语量范畴研究[M]. 武汉：华中师范大学出版社，2000.

[3] 郭先珍. 现代汉语量词用法词典. 北京：语文出版社，2002.

[4] 何 杰. 现代汉语量词研究（修订版）[M]. 北京：民族出版社，2001.

[5] 刘学敏，邓崇谟. 现代汉语名词量词搭配词典.杭州：浙江教育出版社，1989.