新词语自动识别方法研究[1]

郑家恒　李文花

山西大学计算机科学系 030006

摘要：本文在对加工过的网上文本语料统计的基础上，首先用 N 元递增分步算法，获取含新词语的汉字串；经过初筛选，建立新词语候选词库；最后以构词法为依据，对剩余词条进行新词语的识别。封闭测试：召回率为 97%，准确率为 85%以上。

关键字：新词语识别　自动分词　中文信息处理

The Study of Automatic Recognition for New Words

Zheng Jiaheng　Li Wenhua

Computer Science department of Shanxi university 030006

Abstract: This paper uses N-gram increasing algorithm which bases on the analysis of preprocessing text on the web to obtain the character strings containing the new words. After the initial scalping, we constructed the candidate word database of new words. Finally, we recognize the remained words base the word-formation. Close testing, recall rate: 97%, correct rate: higher than 85%.

Keywords: recognition of new words, automatic segment, the processing of Chinese information

1 引言

自古以来科技的进步总是在影响着人类文化的发展，互连网时代的到来，更使这两大阵营迅速地融汇碰撞，出现了奇特的“互连网文化”，由此产生的互连网用语就是其中的代表。

由于网上新词语的出现，使得分词软件对 web 文本的进行分词时，出现过多的“散串”，影响了分词软件的正确率。目前对新词语的识别研究主要集中在新词语的外形特点及外部环境中，而忽略了词本身的内部结构问题。本文在对加工过的网上文本语料统计的基础上，采用 N 元递增分步算法，获取含新词语的汉字串，然后根据汉语构词法，确定新词语。最后给出了相应的实验结果和评价。

2 网上文本中新词语现象分析

2.1 对熟语料中新词语的分析

2.1.1 新词语词长现象分析

[1] 本文得到山西省自然科学基金资助（20001032）

本文所选用的实验语料取自http://sohoo.com.cn中的文本，约10万字，对初加工实验用语料进行了手工抽取新词语，共抽取新词语：二元组86个，三元组57个，四元组42个，具体分析如下：

●双字词语：新词语中的双音节词都集中在连续的单字串上：

例：　只要|网|站|不断|改善|服务|质量　　其中的“网站”；

1|元|左右|的|上|网|成本|实在|让|网|友|羡慕|不|已　其中的“网友”；

●三字词语：新词语中的三音节词，有些是由初加工语料中的双音节词与其相邻的单字构成，也有些是由连续的单字串组成：

例：　网络|股|沉寂|几时　　其中的“网络股”

泡沫|是|新|经济|时代|发展|中|资本|市场|的|产物　　其中的“新经济”

●四字词语：新词语中四音节词出现情况较多，有些由连续的单字构成，有些由相邻的双音节词组成，而有些又是由单音词与其相邻的单字构成：

例：　上|网|资|费|将|继续|下调　　其中的“上网资费”

自打|拥有|了|个人|手|提|电脑|之后　　其中的“手提电脑”

因此，从对新词语词长的组成分析来看，它可以是二元组、三元组、四元组的字串。

2.1.2 新词语构词分析

新词语的构词规则有两种类型：

1.符合常规的构词规则

大部分的新词语的词性结构仍然遵循常有的构词原则：名词与名词、动词、形容词的结合力仍然很强，例：网宅=网（名词）+宅（名词）；网恋=网（名词）+恋（动词）；新经济=新（形容词）+经济（名词）等。

2.特殊的构词规则

由于网络的发展，人类自身对词语的创造性，使得某些特殊词性的字在新词语中有了特殊的意义，从而形成了新的构词规则，如：

①介词具有了构词能力，介词“在”可以和名词结合形成“在线”；

②语气助词“吧”的特殊意义：与另一名词结合，表示一个休息娱乐的场所，形成一个新词，如“网吧”，“水吧”，“冰吧”；

③区别词“黑”原表示一种状态，但在网络却具有了特殊的含义，如：“黑客”、“被黑”；

④在新词语中还出现了一种量词与量词相结合构成新的词语，如：“页面”。

新词语除一部分遵循常规的构词原则外，介词、量词、语气助词等有新的构词能力。

2.2 对初分词后结果的分析

所谓“散串”是指对语料进行初加工后，形成的连续的单字，即切分碎片，连续单字的个数≥2，对实验语料经过分词软件初加工后，成词数为：不同的单字数1446个，不同的二字词5835个、三字词557个，四字词484个，切分碎片为17477次（含重复累计数），其中连续的“散串”共计7916次，占切分碎片的45.3%，单字共计9561次。

通过对切分碎片进行分析，可知，在构成碎片的单字中，1446个单字出现次数从1——4103不等，其中存在着大量的实词部分，其中动词、名词、形容词所占有的比例也很大，在1446个单字中，名词、动词、形容词共计1090个，占总单字数的75.42%。

从上述数据可以看出当分词软件对基于 WEB 的文本进行处理时，出现分词碎片过多，而且碎片中包含了大量的实词部分，我们知道实词具有较强的构词能力，因此有必要对碎片中出现的连续的单字即“散串”进行分析处理。

3 基于 N 元递增的新词语初筛选

3.1 候选词库 H 的获取

定义 1：“N 元递增算法”是指根据新词语词长信息，在对初加工语料进行预处理后，分别从长度 N=1、N=2、N=3 的词串中获取二元组、三元组、四元组的字串。

下面以二元组候选词库的形成为例说明“N 元递增算法”。

① 先将二字词的候选词条集 twowordtable 置空，然后对基于 Web 的网上语料进行初加工，将初加工文档中的 N 元词（N≥2）及数字、西文字符等一切非汉字字符去掉均以空格代替；

② 剩余的文档可以表示成一个含有许多空格和汉字的字符串 getstring；从 getstring 的首端开始向尾部进行扫描，以连续的 2 个汉字作为匹配字串，查找 twowordtable，如果找到匹配的模式，则将字串的频度加 1，如果没有，则将该新字串加到 twowordtable；

③ 如此重复上述扫描过程，一直到 getstring 被搜索完毕，获得每一种汉字串的结合模式，并统计出该模式在文档中出现的频度及汉字字串的结合方式。

3.2 候选词库 H 的过滤

3.3.1“功能字”、“功能词”的剔除

定义 2：所谓“噪声字串”是指在候选词库中具有明显特征的，不具备构词能力的汉字结合模式。

定义 3：功能字是指那些构词能力弱，且在语料中出现频次高的单字，例如：“我”、“的”“啊”、“啥”等。

定义 4：功能词是指那些长度为 2 的虚词，包括：代词、连词、介词、副词、语气词、助词等，如：“我们”、“可是”、“朝着”、“本来”、“也罢”、“等等”等。

N 元组的获得是对初加工语料中符合候选词条出现位置的连续的汉字串利用 N 值递增算法得到的，由于在经过初加工的语料中，“散串”中含有大量的“功能字”与“功能词”，因此必须首先利用“功能字”、“功能词”的剔除原则，将含有“功能字”与“功能词”的“噪声字串”去掉。例如：

|手机|N|及|C|宽|N|带|N|网络|N|是|V|最|D|炙手可热|I|的|U|话题|N|

经过 N 值递增算法，形成如下字串：

二元组	及宽　宽带　是最
三元组	手机及　机及宽　及宽带　宽带网　带网络　网络是　络是最　的话题
四元组	手机及宽　机及宽带　及宽带网　宽带网络　带网络是　网络是最

表 3.1　候选字串表

由此可见，形成N元组共17个，经过“功能字”与“功能词”去除后，表3.1所形成的N元组情况如下：

二元组：宽带　　三元组：宽带网　带网络　　四元组：宽带网络

3.3.2“可存在性”的过滤

定义5：所谓“可存在性”是指候选词库中的字串是否可以作为候选词条存在的依据，以该字串出现的频度作为“可存在性”的度量值。

由N元递增算法可知：在形成N元新词的过程中，由于句子中连续汉字的出现而形成很多“偶然型”的汉字串，例如：“带网络”，通过对N元组候选词库的观察可知，这些“偶然型”噪声字串的频度大多为1和2。

性质1：候选词库中的字串可作为侯选词条存在的依据是该汉字串的频度必须大于1。

经过“可存在性”过滤，剩余字串如下：

二元组：宽带　　三元组 ：宽带网　　四元组：宽带网络

3.3.3“N元重叠”的过滤

定义5：由于对不同的N值，其N元统计是分别计算的，因此文本中一个给定字串的每一次出现都可能作为不同N值的若干N元组的一部分而被计算，例如：“手提电脑”的出现亦被“手提电”、“提电脑”统计，这种现象称为N元重叠。

对N元重叠的过滤采用频率相减法，长字串N元组Y的频率f（Y）-短字串的N元组X的频率f（X），建立规则：

R1：前提条件：　若差值≥0，结论：说明X每次均出现在Y中，则将其去除。

R2：前提条件：若差值<0，结论：则说明X亦作为N元组单独出现，则将X保留，且f（X）=f（X）-f （Y）。

R3：前提条件：若差值＜0且f（Y）≤3，结论：则将长字串N元组Y去除。

如：PL（手提电脑）=31，f（手提电）=31，f（提电脑）=31

因为：　f（手提电）=f（手提电脑）-f（手提电）=0

f（提电脑）=f（手提电脑）-f（提电脑）=0

因此，进行N元重叠过滤后，“手提电”和“提电脑”被过滤掉。

4 基于构词规则的新词语确定

运用构词法知识，建立构词规则库，规则约32条，分为三种类型。

4．1“互斥性字串”的过滤规则库

定义6：所谓“互斥性”是指在候选词库中，构成词条的各组成部分的词性不符合构词规则，将其去除。

规则1：前提条件：若字串中的第一个字是副词，结论：将该字串从候选词库中去除。

规则2：前提条件：若字串中的含有连词，结论：将该字串从候选词库中去除。

4．2 常规构词规则库

规则 1： 前提条件：若两个单字词均为名词，结论：将该候选词条读入新词库。

规则 2： 前提条件：若候选词条最后一个字为后接成份，且前两个为名词性双音节词，结论：将该候选词条读入新词库。

规则 3：前提条件：若候选词条由两个名词性的双音词组成，结论：将该候选词条读入新词库。

4．3 特殊构词规则库

规则 1：前提条件：若候选词条的第二个字是“吧”，且第一个单字词为名词，结论：将该候选词条读入新词库。

规则 2：前提条件：若候选词条的前两个字为双音节词“网络”，且其后接名词性双音节词，结论：将该候选词条读入新词库。

5 实验及分析

实验语料取自 http：//www.sohu.com 中的网络传情栏目，主要包括网民对网络发展的评述、对网络现状的探讨及网民所写的随想杂文，语料库规模 10 万字，经测试新词语的召回率为：97%，准确率为：85%。

5.1 新词语自动识别的系统结构图：

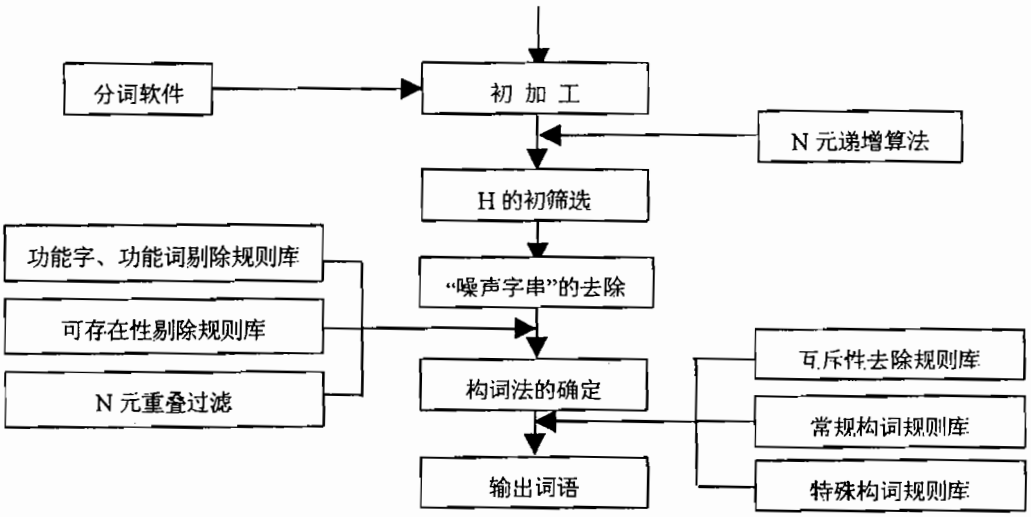

图 5—1 新词语自动识别系统框图

5.2 实验测试数据与结果分析

对封闭实验语料进行新词语识别，测试数据分布如下：

	二元组数目	三元组数目	四元组数目
形成N元组数	6675	47248	50664
“噪声字串”过滤后	998	1363	587
构词规则库的运用	128	235	142
形成的新词语数	98	70	55

表5—1 实验数据分析

从以上数据可以看出，在进行“噪声字串”剔除后，二元组的剔除率在85%以上，而三元组与四元组的剔除率在90%以上，为什么会有如此高的剔除率呢？这是因为新词本身存在于初加工语料的“散串”或部分分词单位的组合而成，根据《信息处理用现代汉语分词规范》可知，在语料进行初加工后，之所形成“散串”，就因其是单个的助词、连词、副词、介词、语气词等分词单位，即所谓的“功能字”，同时也含有大量的双音节的连词、副词、介词、助词、叹词等分词单位，即所谓的“功能词”，这种剔除是安全的，由于在新词语中存在有特殊的构词规则，对于误剔除的词可以采用特殊规则库中的规则进行召回。

6 结束语

未登录词的自动识别是自动分词中的一个瓶颈问题。不可能将所有的未登录词都加入词典，因此，有必要采用一种算法，对使用分词软件进行语料加工而形成“散串”较多的部分进行新词语召回。

基于规则剔除与构词法相结合的新词语识别技术，利用初加工语料，采用分解策略将N元组候选词库的形成分为预处理、二元候选字段、三元候选字段、四元候选字段几个过程，降低了整体处理难度，使得候选词库中包含了所有可能出现的新词，虽然也包含大量“噪声字串”，但是通过运用“功能字”、“功能词”、“可存在性”及结合字串词性的互斥性，使“噪声字串”的对新词语的识别所造成的影响降到最低，再充分地利用了汉字（词）的自身属性特点与构词法相结合，对新词语进行召回，取得了很好的效果，是一种可行的办法。

参考文献

[1]刘开瑛, 中文文本自动分词和标注, 商务印书馆，2000年北京

[2]Jian-yun Nie,Marie-Louise Hannan,etc. “Unknown Word Detection and Segmentation of Chinese using Statistical and Heuristic Knowledge”,Communications of COLIPS,vol.5,No 1&2,DEC 1995,69-77

[3]郑家恒、李鑫、谭红叶，基于语料库的中文性名识别方法与研究，中文信息学报，2000，14

[4] 苑春法、黄昌宁，基于语素数据库的汉语语素及构词研究，世界汉语教学，1998年第2期

[5]陆志苇,《现代汉语构词法》(修订本)，中华书局，1975年