# 基于新词发现的古典文学作品分词方法研究

高嘉琦[1],赵庆聪[1,2]
(1. 北京信息科技大学 信息管理学院,北京 100192;
2. 绿色发展大数据决策北京市重点实验室,北京 100192)

**摘 要**:对于中文文本的分词研究来说,现有的分词方法和技术较多都是针对现代汉语,现代汉语的分词方法和体系已经很成熟,但对古代汉语的研究较少。由于古文的特殊性,将现代汉语的分词方法技术直接用于古汉语时,无法得到分词准确的理想效果,目前对古汉语分词方法的研究还未形成成熟的体系。文中提出一种基于新词发现的古典文学作品分词方法,即从大量古典文学作品语料中发现新词,构建古汉语分词词典,在此基础上再对古文文本进行分词。以《三国演义》古文文本处理为例,验证了基于新词发现的古典文学作品分词方法能有效提高古文分词的准确率。

**关键词**:古典文学;新词发现;分词;互信息;左右熵
**中图分类号**:TP301 **文献标识码**:A **文章编号**:1673-629X(2021)09-0178-04
doi:10.3969/j.issn.1673-629X.2021.09.030

# Study on Word Segmentation Method of Classical Literature Based on New Word Discovery

GAO Jia-qi[1], ZHAO Qing-cong[1,2]
(1. School of Information Management, Beijing Information Science and Technology University, Beijing 100192, China;
2. Beijing Key Laboratory of Big Data Decision for Green Development, Beijing 100192, China)

**Abstract**: For the research on word segmentation of Chinese text, most of the existing word segmentation methods and technologies are aimed at modern Chinese. The word segmentation methods and systems of modern Chinese have been quite mature, but there are few studies on ancient Chinese. Due to the particularity of ancient Chinese, when the modern Chinese word segmentation method and technology are directly used in ancient Chinese, the ideal effect of accurate word segmentation cannot be obtained. At present, the word segmentation of ancient Chinese has not yet formed a general method and mature system. We propose a method of word segmentation in classical literature based on neologism discovery, that is, discovering new words from a large number of classical literary works, constructing an ancient Chinese word segmentation dictionary, and then segmenting the ancient text on this basis. Taking the ancient text processing of "The Romance of the Three Kingdoms" as an example, it is verified that the word segmentation method of classical literary works based on the discovery of new words can effectively improve the accuracy of ancient text segmentation.

**Key words**: classical literature; new word discovery; word segmentation; mutual information; left-right entropy

## 0 引 言

中国历史文化源远流长,有大量的古典文学作品流传至今,这些作品展现了中华民族从古至今的人文精神,同时也传承了上下五千年以来优秀的传统价值观,是一笔宝贵财富。对古典文学作品进行深入的分析和研究,在文化传承、历史研究、人文教育等方面都有十分重要的意义。

对古典文学作品进行文本挖掘,分词是基础。目前现代汉语的分词方法技术已较为成熟,而对古代汉语的分词处理尚处于探索、验证阶段。国内学者对古文分词已进行的研究有:石民[1]等采用条件随机场模型,基于两个模板进行古文分词标注一体化,使得准确率和召回率都有所提升。王嘉灵[2]选用条件随机场模型,以《汉书》为语料,并选择了核实的特征模板,制定了《汉书》分词规范,进行分词实验,实验结果的F值达到94.4%。王晓玉等[3]选用中古时期的语料,选用条件随机场和词典的方法训练分词模型,解决了人工分词不一致问题。杨世超等[4]采用带有古汉语特征的条件随机场模型作为特征模型,获得了较好的分词效果。

**收稿日期**:2020-09-23 **修回日期**:2021-01-25
**基金项目**:国家重点研发计划项目(2017YFB1400400)
**作者简介**:高嘉琦(1994-),女,硕士研究生,研究方向为文本挖掘;赵庆聪,博士,副教授,研究方向为数据分析。

上述研究都需要大量的人工标注,费时费力,缺乏通用性,也未能提出一种能快速构建古汉语词库的有效方法[5]。古典文学作品中有大量的词汇已不在现代使用,故也未收录到现代汉语词典中,这是造成分词效果差的主要原因,如果对古典文学作品进行新词发现,构建古汉语分词词典,能有效提高分词的准确率。目前,中文新词发现的研究主要集中在现代文语料,由于古文在字词、短语和语法结构方面都与现代文有所不同,所以,现有的现代文语料上的新词发现技术无法直接应用于古文语料[6]。文中提出一种基于新词发现的古典文学作品分词方法。首先,对古典文学作品采用 N-Gram 算法进行切分,然后采用互信息和左右信息熵的新词发现方法识别新词,将新词发现识别出的新词与原有的基础词典相结合,构建出古文分词词典,再使用 Jieba 中文分词器对古典文学作品进行分词,最后通过实验以检验分词的准确度。

## 1　新词发现的相关技术

基于规则的新词发现方法、基于统计的新词发现方法和基于统计与规则相结合的新词发现方法是现有常用的新词发现方法[7]。基于规则的新词发现方法[7-8]是指使用词语的特性和成词的原理和语义的特征来构建数学模型对文本中的新词进行挖掘。该方法具有较高的准确性,但具有较差的可扩展性、通用性,后期维护也困难,需要人工构建规则库,会消耗大量的人力和物力,无法满足新词出现速度快、消亡快的需求。基于统计的新词发现方法[8-10]是指通过大量的实验对文本语料进行处理,计算词语的词频、成词的概率、左右邻接熵、邻接变化数等统计特征来识别新词。这种新词发现方法有较强的普适性,方便扩展和移植,不受不同种类文本的限制,但需要对模型进行大量训练,同时具有准确率较低的缺点。基于规则与统计相结合的新词发现方法是尽量将两种方法的优点相结合,从而使新词发现方法更加准确也更高效[7]。

文中先采用 N-Gram 算法切分古文语料,得出候选词集,再采用规则与统计相结合的新词发现方法,即互信息、左右信息熵的统计特征与停用词、过滤首尾停用词等规则相结合,最终实现新词发现。

### 1.1　N-Gram 算法

N-Gram 是一种基于统计语言模型的算法,用于切分语料得出候选词集,方便后续计算词语的内部凝固度和自由程度。N-Gram 算法的具体思路是:使用大小为 $N$ 的滑动窗口对文本语料按字节流进行滑动操作,形成每个字节的片段称为 gram,形成的片段是长度为 $N$ 的字节片段序列,提前设定阈值对 gram 按照出现的频度进行过滤,形成关键 gram 列表,列表中的每一种 gram 均为一个特征向量维度[11]。一般情况下,取 $N=3$ 的情况较多。如果 $N$ 的取值太大,会造成等价类太多,自由参数过多。

### 1.2　互信息

在信息论相关领域中,互信息(mutual information)是指两个事件集合之间的相关性,是一种有用的信息度量[12]。互信息度量的是两个随机变量之间的统计相关性,是从随机变量整体角度,在平均的意义上观察问题,因此通常称之为平均互信息。互信息表示两个变量或多个变量之间共享的信息量,互信息越大,变量之间的相关性越强[13]。在文中,词语是文章的最小结构形式,可以独立存在,词语中的相邻的字之间都有一定的关联性。如果词语中字与字的这种关联性越大,说明可能是词的可能性也就越大。可以用互信息计算新词的内部成词概率,互信息一般可用于表示两个事件相互关联的程度,互信息值越大,表示两个物体的关联程度也就越大。在词汇聚类、汉语自动分词、词义消歧、文本分类和聚类等问题的研究中,互信息也具有重要用途。互信息用以下公式来计算:

$$\mathrm{PMI}(m,n)=\log_2\frac{p(m,n)}{p(m)p(n)} \tag{1}$$

其中,$p(m)$ 表示字符 $m$ 单独出现在语料集中的概率;$p(n)$ 表示字符 $n$ 单独出现在语料集中的概率;$p(m,n)$ 表示字符 $m$ 和字符 $n$ 组合起来共同出现在语料集中的概率;$\mathrm{PMI}(m,n)$ 表示字符 $m$ 和字符 $n$ 的相互关联程度。若 $\mathrm{PMI}(m,n)>0$,表示字符 $m$ 和字符 $n$ 是相互关联的,而且 PMI 的值越大,表示两者相互关联的程度越大,也就越有可能成为新词;若 $\mathrm{PMI}(m,n)=0$,则表示字符 $m$ 和字符 $n$ 是彼此独立的。

### 1.3　左右信息熵

熵是信息论的基本概念。熵又称为自信息,熵可以作为数量用来描述一个随机变量的不确定性。若用来描述随机变量的熵越大,那这个随机变量的不确定性越大,越不确定的随机变量越需要大的信息量用以确定其值,正确估计其值的可能性就越小。信息的作用是消除人们对事物的不确定性,信息熵是对信息的量化度量,信息熵值越大则事物的不确定性也越大,所需要的信息量也就越大。候选新词的左边邻接词和右边邻接词的不确定性可以用左右信息熵来衡量,其不确定性越大,说明该词的周边词越丰富,其成词的概率就越高。左信息熵和右信息熵的计算公式为:

$$E(\mathrm{prew})=-\mathrm{sum}(p(\mathrm{prew})\log_2 p(\mathrm{prew})) \tag{2}$$

其中,prew 是候选词邻接字的集合,$p(\mathrm{prew})$ 表示候选词的左右邻接字的条件概率。

## 2　基于新词发现的古典文学作品分词方法

对古典文学作品进行分词是对古典文学作品进行

研究的基础。基于词表的分词方法和基于统计的分词方法是目前古汉语的自动分词任务常用的方法[14]。基于词表的分词方法需要人工标注词汇构建古籍文本词典,通过古籍文本词典进行分词[5]。这种分词方法准确率较高,但要耗费大量的人力物力,具有局限性;基于统计的分词方法需要训练人工标注的分词语料,使用学习模型,从而实现古籍文本自动分词。以上两种方法都需要先进行人工标注训练集,人工标注需要较高的专业知识,而且需要大量时间,难度和成本都比较高。

### 2.1 古典文学作品分词方法

文中首先使用 N-gram 算法对古文语料进行切分,统计各个词的词频,使用词频和过滤停用词等相关规则进行初步筛选,得到初始词表;然后用互信息计算内部凝固度来对词表进行第二次筛选;最后用左右信息熵对二次筛选后的词语计算其自由程度,根据自由度值进行再次筛选,最终确定新词词表。将获得的新词词表添加到 Jieba 中文分词器中,形成古文分词词典,再对古典文学作品进行分词。这种方法省去了人工标注环节,可快速构建古文分词词典。分词流程如图 1 所示。

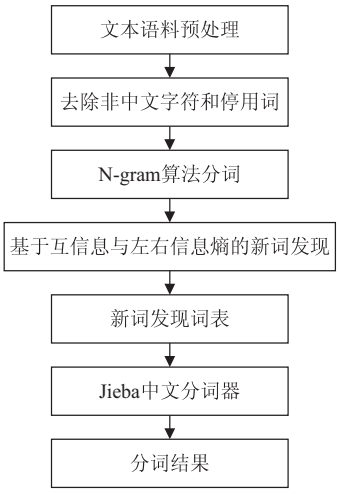


图 1 基于新词发现的古典文学作品分词流程

### 2.2 分词方法的具体实现

本研究选用了经典的文言文章回小说作为文本语料。著名语言文学家王力先生在《古代汉语》中指出:“文言是指以先秦口语为基础而形成的上古汉语书面语言以及后来历代作家仿古的作品中的语言”[15]。顾名思义,文言文就是用文言写成的文章,是中国古代的书面语言,沿用了两三千年,也是现代汉语的源头[15]。文言文章回小说篇幅长,既有古文的结构和语法特点,又有相当数量的词汇沿用到现代,便于研究人员采用现代文的分词词库作为基础词典,在此基础上进行新词发现。

文中选择了包括《三国演义》、《聊斋志异》、《镜花缘》、《说唐》等在内的 68 部章回小说文本作为基本语料,经统计有 27 960 539 个汉字。

(1)古文文本预处理。将文本转换为 TXT 格式,利用正则表达式过滤非中文符号——将古文文本中用于断句的标点符号、特殊符号等噪声数据过滤掉,得到预处理之后的文本语料。

(2)语料切分。使用 N-gram 算法对预处理过的文本语料从左至右逐字进行切分。由于古典文学作品中有三字词语,如人名等。设置 $N$ 为 3,并得到 1-gram ~3-gram 包含词频的 gram 词表,获得初始候选新词结果。

(3)计算候选词的互信息。先将单字过滤掉,然后对其余初始候选新词计算互信息,若该词语的互信息大于设置的阈值,将其保留,生成候选新词集。

(4)计算候选词的左右信息熵。对候选词进行左信息熵和右信息熵的统计,将左信息熵和右信息熵相加,得到左右信息熵。设置左右信息熵的阈值,若该词的左右信息熵大于设置的阈值,将其保留,最终得到新词集。

算法流程如图 2 所示。

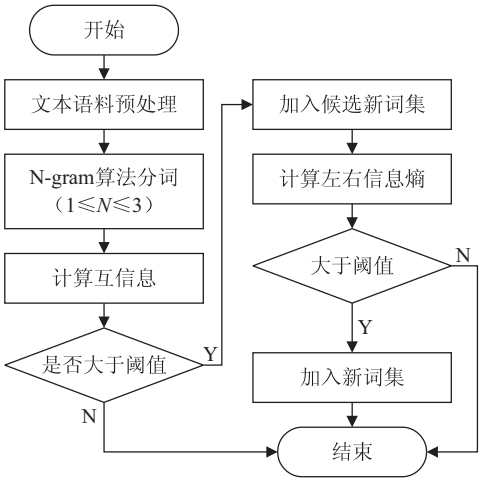


图 2 算法流程

## 3 实验及结果分析

### 3.1 实验语料

文中以文言文章回小说《三国演义》为例,展示使用基于新词发现的古典文学作品分词方法的分词结果,并对分词效果进行了分析。

对整篇《三国演义》文本语料进行预处理后,使用 N-Gram 算法对文本进行切分,切分部分结果如图 3 所示。

对上述切分得到的初始候选新词计算互信息,互信息值大于设置阈值的保留,生成候选新词集,得到 16 081 个候选新词。

再利用左右信息熵的算法进行筛选,得到最终的

新词集合,获得 3 892 个新词,部分新词结果如图 4 所示。

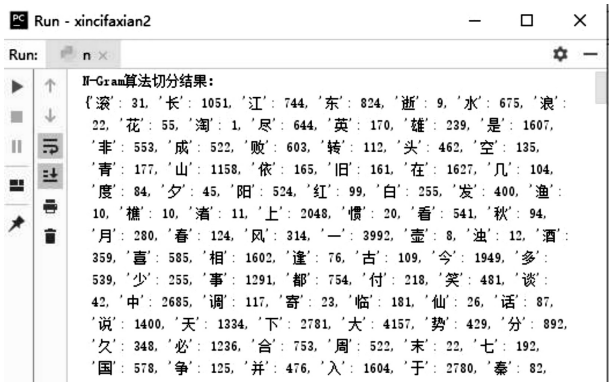


图 3　N-Gram 算法的部分切分结果

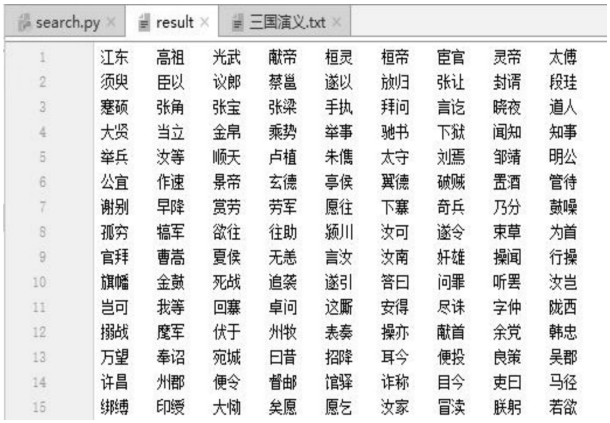


图 4　新词发现的部分结果

以《三国演义》第九十八回中的部分语料为例,从中提取了 3 个新词:孔明、蜀兵、魏兵。

在加入新词前使用 Jieba 中文分词器进行分词结果为:所忧者/但/魏延/一军,在/陈仓道/口/拒住/王双,急/不能/脱身;吾/已/令人/授以/密计,教/斩/王双,使/魏人/不敢/来/追。蜀/兵/更/不/回头。双/拍马/赶来。背后/魏/兵/叫/曰:“城外/寨中/火/起,恐/中/敌人/奸计。”后人/有/诗/赞曰:“孔明妙/算/胜/孙/庞,耿若长/星/照/一方。进退/行/兵/神/莫测,陈仓/道口/斩/王双。”[16]

加入新词之后,Jieba 中文器的分词结果为:所忧者/但/魏延/一军,在/陈仓道/口/拒住/王双,急/不能/脱身;吾/已/令人/授以/密计,教/斩/王双,使/魏人/不敢/来/追。蜀兵/更/不/回头。双/拍马/赶来。背后/魏兵/叫/曰:“城外/寨中/火/起,恐/中/敌人/奸计。”后人/有/诗赞曰:“孔明/妙算/胜/孙庞,耿若长/星/照/一方。进退/行/兵/神/莫测,陈仓/道口/斩/王双。”[16]

### 3.2　评价指标

文中采用准确率 $P$(precision)、召回率 $R$(recall)和 $F$ 值(F-measure)作为评价指标,来检验利用基于互信息与左右信息熵的新词方法发现的实验结果,计算公式如下:

$$P = \frac{N \cap M}{N} \times 100\% \tag{3}$$

$$R = \frac{N \cap M}{M} \times 100\% \tag{4}$$

$$F = \frac{2PR}{P + R} \tag{5}$$

其中,$N$ 表示实验获得的新词的总数;$M$ 表示古典文学作品中本身存在的新词总数($M$ 值为经古汉语专家人工标注的新词数量)。

利用文中方法对《三国演义》进行新词发现,得到的结果如表 1 所示。

表 1　新词发现评价结果

| 实验名称 | 准确率/% | 召回率/% | $F$ 值 |
|---|---|---|---|
| 互信息与左右信息熵的新词发现 | 58.02 | 33.63 | 42.58 |

结合表 1 和对比分词结果,虽然新词发现的准确率、召回率和 $F$ 值略低,但通过比较加入新词前后的两个分词结果,加入新词之后分词的准确度有明显提高。

## 4　结束语

文中采用互信息和左右信息熵的新词发现方法对古典文学作品挖掘未登入的新词,利用 Jieba 中文分词器结合新词词表,对古典章回小说进行分词实验,分词效果得到明显改善。该方法避免了古汉语文本分词需要大量人工标注的问题,快速构建了古汉语分词词典,为后续对古典文学作品的深入研究打下了坚实的基础。该方法的不足之处是新词发现的准确率、召回率、$F$ 值都不高,未来还需要进一步研究,以提高新词发现和分词的准确率。

**参考文献:**

[1] 石　民. 基于 CRF 的古汉语分词标注一体化研究[C]// 中国计算机语言学研究前沿进展(2007-2009). 烟台:中国中文信息学会,2009:6.

[2] 王嘉灵. 以《汉书》为例的中古汉语自动分词[D]. 南京:南京师范大学,2014.

[3] 王晓玉,李　斌. 基于 CRFs 和词典信息的中古汉语自动分词[J]. 数据分析与知识发现,2017(5):62-70.

[4] 杨世超,纪　月,赵立鹏. 基于条件随机场的古汉语分词研究[J]. 电脑知识与技术,2017,13(22):183-184.

[5] 李筱瑜. 基于新词发现与词典信息的古籍文本分词研究[J]. 软件导刊,2019,18(4):60-63.

[6] 谢　韬. 基于古文学的命名实体识别的研究与实现[D]. 北京:北京邮电大学,2018.

[7] 刘伟童,刘培玉,刘文锋,等. 基于互信息和邻接熵的新词发现算法[J]. 计算机应用研究,2019,36(5):1293-1296.

(下转第 207 页)