计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(5):927~932, 2006

基于大规模语料库的新词检测

崔世起[1,3] 刘 群[1] 孟 遥[2] 于 浩[2] 西野文人[2]
[1](中国科学院计算技术研究所数字化技术研究室 北京 100080)
[2](富士通研究开发中心有限公司 北京 100016)
[3](中国科学院研究生院 北京 100049)
(sqcui@ict.ac.cn)

New Word Detection Based on Large-Scale Corpus

Cui Shiqi[1,3], Liu Qun[1], Meng Yao[2], Yu Hao[2], and Nishino Fumihito[2]
[1](*Digital Technology Laboratory*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100080)
[2](*Fujitsu Research & Development Center Co.*, *LTD*, *Beijing* 100016)
[3](*Graduate University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract New word detection is a part of unknown word detection. The development of natural languages requires us to detect new words as soon as possible. In this paper, a new approach to detect new words based on large-scale corpus is presented. It first segments the corpus from the Internet with ICTCLAS, and searches for repeated strings, and then designs different filtering mechanisms to separate the true new words from the garbage strings, using rich features of various new word patterns. While getting rid of the garbage strings, three garbage lexicons and a suffix lexicon are used, which are learned by the system, and good results are achieved. Finally, the results of the experiments are discussed, which seem to be promising.

Key words new word; garbage string; garbage head; garbage tail; IWP

摘 要 自然语言的发展提出了快速跟踪新词的要求.提出了一种基于大规模语料库的新词检测方法,首先在大规模的 Internet 生语料上进行中文词法切分,然后在分词的基础上进行频度统计得到大量的候选新词.针对二元新词、三元新词、四元新词等的常见模式,用自学习的方法产生 3 个垃圾词典和一个词缀词典对候选新词进行垃圾过滤,最后使用词性过滤规则和独立词概率技术进一步过滤.据此实现了一个基于 Internet 的进行在线新词检测的系统,并取得了令人满意的性能.系统已经可以应用到新词检测、术语库建立、热点命名实体统计和词典编纂等领域.

关键词 新词;垃圾串;垃圾头;垃圾尾;独立词概率

中图法分类号 TP391

1 引 言

汉语通过派生、复合、缩写等形式产生了很强的造词功能[1],而任何一部汉语词典都不可能包含所有的中文词.在词典中不存在的词称为未登录词,即未登录词的外延是由使用的词典决定的.未登录词识别是中文信息处理中的难点和热点,对中文词法切分发挥着重要作用.中文词法切分的大部分错误是由未登录词的识别错误造成的.未登录词主要的形式包括人名、地名、机构名等命名实体,以及时间词、数量词和普通的语法派生词等[2].对命名实体的研究[3~5]较多,而且已经取得了较好的效果,但对普通新词的识别缺乏广泛的研究.但随着政治、经济、文化的进步,人类生活方式的革新,自然语言中出现了大量的新词.大规模的搜集新词来扩充

收稿日期:2005-03-04;修回日期:2005-08-11
基金项目:国家"八六三"高技术研究发展计划基金项目(2004AA114010,2003AA111010);中国科学院计算技术研究所和富士通研究开发中心有限公司合作项目

现有的词典,是一项很有意义的工作;而把新词检测方法应用到中文词法分析中,也对分词性能的提高很有帮助.本文的讨论重点是对普通新词的检测.

我们把研究的新词定义为未登录词中除去命名实体、时间词和数量词之外的普通的语法派生词.根据新词的构成方式,本文研究的新词分为4类.

缩写:非典(非典型肺炎)、边警(边防警察)、抗非(抗击非典).缩写词的构词方式很不规则,使用词中的某个字表示词的含义,但该字的选择有时候只是一种约定或者习惯,所以缩写词的检测是很难的.

派生词:垂直化、价值型.这类词有比较明显的词缀语素.

复合词:扑杀、现金流.通过复合会产生大量的新词,复合方式多种多样,汉语中很多活跃的字都可能作为复合词的元素,这类词的识别也是非常难的.

单纯词:肯德基、麦当劳.该类词的意义与单字的意义完全无关,包括音译词等.

新词是未登录词的一种.已有的对未登录词的检测方法研究重点是命名实体.命名实体具有明显的标志,而且词性为名词,在句子中担任着固定的句法功能,检测起来相对容易.新词构词形式多样化,而且可能是名词、动词、形容词或者副词等,句法功能并不固定,单纯从新词的内部结构去考虑很难做出准确的判断.

下文主要从以下几个方面来进行论述:①相关工作,介绍已有的新词检测的研究成果.②基于大规模语料库的新词检测方法,详细介绍本文提出的新词检测方法的过程.③性能分析,通过实验对该方法进行分析.④结论和下一步工作.

2 相关工作

新词的检测方法有两种:一种是基于规则的方法.郑家恒等人[6]根据汉语构词法建立规则库,通过调用'互斥性字串'过滤规则和构词规则来进行网络新词语的识别.第2种是基于统计的方法.IBM专利中首先统计串频,而后对结果进行修剪.Wu等人[7]提出IWP(独立词概率)对被切散为单字串的新词进行识别.条件随机域[8](conditional random fielsd)模型被用到中文词法切分中,并通过对切分片断计算置信度来寻找可能的新词(Peng等人).Min-Jer Lee等人[9]在live dictionary的研究中,利用在线的文本资源自动获取专业术语和相似术语等新词来构建可以动态增长的词典.邹纲等人[10]提出了一种以时间参数为判断依据的面向Internet的新词检测方法,根据重复串的出现时间和频度来进行新词识别.

使用基于规则的方法,新词检测的准确率都较高,但如果人工提取规则,会消耗大量的人力,而且规则的覆盖性不会太好,随着新词的不断产生,需要不断地添加规则.如果自动提取规则,规则的有效性得不到保证.使用统计的方法,利用频度信息来确定新词,会引入许多频度较高的垃圾串.本文提出一种基于大规模语料库的方法,不仅利用了新词的词频信息,而且通过分析新词和垃圾串不同模式的特征,解决了垃圾串过滤的问题,使得系统取得了较高的准确率和召回率.

3 基于大规模语料库的新词检测方法

3.1 理论前提

通用度和使用度是评价词语的两个参数.《现代汉语通用词——基本集》中提出了兼顾频率因素和分布因素,提出了"通用度"的计算公式,并根据词语的通用度进一步把词分为4个级别.Juillan和Chang-Rodsiguez在计算西班牙语的词汇频率时提出使用度的概念,可以综合地反映单词在出现频率和分布率两方面的情况.

我们希望检测到的新词是通用度和使用度比较高的词,因为这种被频繁使用的新词更能反映社会发展的趋势,更可能为大众所接受.为了研究新词的出现频率和分布率这两个重要指标,我们提出了基于大规模语料库的新词检测方法.因特网就是一个庞大的语料库,从这样的真实语料中采集的网页是最能反映语言现象和新词分布特征的.我们使用总频率表示新词的出现频率特征,用文档频率表示新词的分布特征.

图1是新词检测方法的过程图.

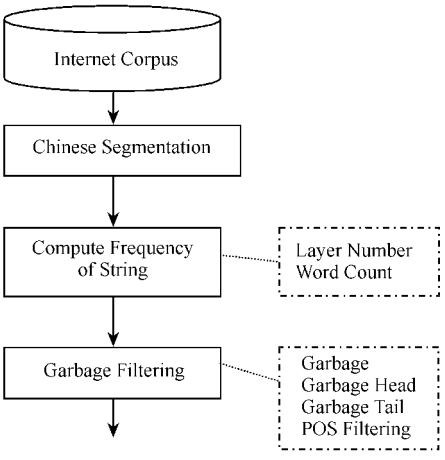

Fig. 1 Procedure diagram of new words detection.

图1 新词检测过程图

3.2 语料库收集和串频统计

第1步就是获取大规模的真实语料.我们在系统中实现了在批量Web网页中抽取文本的功能.

第2步是串频统计.纯粹的基于字的统计会产生过多的不符合语法规则的垃圾串.为减少垃圾串的数量,我们首先使用ICTCLAS对文本进行切分标注,然后在词的基础上进行重复串搜索.

串频统计具有以下优点:①对构词方式比较新颖的新词有较好的召回率.像"非典"、"虐俘"等新词利用规则方法容易被过滤掉.②对外来词和方言词有较好的召回率.③对较长复合词的判断提供了一种依据.有些复合词,尤其是词+词模式的组合,到底应认为是短语还是新词很难定论.我们可以根据词频来判断,那些比较高频的组合就可以认为是新词.④通过设置较高的串频阈值,我们可以得到通用度和流通度较高的新词.

3.3 垃圾串过滤

我们分析新词可能的形态发现,通过词法切分,新词被切成了碎片.KehJiann Chen对Sinica Corpus的部分语料统计得出结论,4632个未登录词中有4372个包含至少一个单字词,而完全由多字词组成的新词只有60个.因此对2字、3字和4字新词来说,我们主要对以下模式的新词进行检测:1+1模式、1+1+1模式、1+1+1+1模式、2+1模式、1+2模式、1+2+1模式、3+1模式和1+3模式.举例如下:

他们 诉求 工资 待遇 的 提高 、工作 环境 的 改变 (1+1模式).

湖北 农民工 讨薪 坠楼 身亡 (1+1模式).

福建 综合 防治 初步 遏制 偷私渡 高发 态势 (1+1+1模式).

实际 上 警告 台 不要 搞" 统独公投 " (1+1+1+1模式).

日军 遗弃 化学 毒剂 伤人 事件 善后款 正式 发放 (2+1模式).

黄昌宁在95万词的语料中统计得出,2+1模式的新词大约是1+2模式的新词的10倍.我们在千龙网站随机下载的1500张网页内,以一张网页为搜索单位,对各种模式的串进行频度统计发现,三元词中2+1模式的新词占绝大多数,四元词中3+1模式的新词占绝大多数,而且由于1+2模式、1+2+1模式和1+3模式的新词内部结构没有明显的特征,识别准确率很低,我们不把它们作为新词检测的重点.

对垃圾串的处理,也就转化为对以下模式的垃圾串的过滤.分析切分的结果,我们发现垃圾串主要有以下形式:

他 在 视察 工作 (1+1模式).

台风 马上 就 要 在 浙江 登陆 (1+1+1模式).

该 谷 的 一片 平地 即 被 辟 为 马场 (1+1+1+1模式).

从《 中学生 不必 读 》看 中学生 读书 现状 (2+1模式).

希丁克 式 的 管理 (3+1模式).

在对切分结果进行频度统计之后,会得到大量如上所示的垃圾串.能否有效地过滤垃圾串,是新词检测的关键.针对单字串(即1+1、1+1+1和1+1+1+1模式的串)和2+1、3+1模式的垃圾串,我们提出不同的方法来进行过滤.

3.3.1 单字串的过滤机制

汉语中单字垃圾串多由一些介词、副词、连词和一些单字实词产生.我们可以在切分好的语料库上训练,获取常用单字串的集合.我们使用了《人民日报》1998年1月到6月切分标注好的语料库来训练,覆盖了绝大多数的常用单字串,以此对垃圾串进行过滤.对于新词伴生的垃圾串,我们使用学习到的垃圾头词典和垃圾尾词典进行过滤,取得了满意的效果.

算法过程:

① 寻找训练语料中所有的单字串碎片,每一个碎片作为一个词典项添加到垃圾词典Garbage Lexicon中.

② 对于Garbage Lexicon中的任意字,计算它作为垃圾词典项的首字的概率.如果概率大于某一个阈值,我们把该字添加到垃圾头词典Garbage Head Lexicon中.

③ 对于Garbage Lexicon中的任意字,计算它作为垃圾词典项的尾字的概率.如果概率大于某一个阈值,我们把该字添加到垃圾尾词典Garbage Tail Lexicon中.

④ 对于任意一个候选新词 W,生成它的所有两个字以上的子串 $W_1, W_2, \cdots, W_n$.

⑤ 如果存在 W_i $(1 \leqslant i \leqslant n)$ 出现在垃圾词典Garbage Lexicon中,则 W 非新词,若 W 去掉 W_i 后仍为单字串碎片,则把该碎片加入到候选新词集合中,转到④处理下一个候选新词;否则转到下一步.

⑥ 如果 W 的首字 BH 出现在垃圾头词典

Garbage Head Lexicon 中,则 W 非新词,若 W 去掉 BH 后仍为单字串碎片,则把该碎片加入到候选新词集合中,转到④处理下一个候选新词;否则转到下一步.

⑦ 如果 W 的尾字 BT 出现在垃圾尾词典 Garbage Tail Lexicon 中,则 W 非新词,若 W 去掉 BT 后仍为单字串碎片,则把该碎片加入到候选新词集合中,转到④处理下一个候选新词;否则 W 通过垃圾串检测.

3.3.2 2+1,3+1 模式串的过滤机制

2+1,3+1 模式的新词有比较显著的特征:尾字常为组合性强的非词语素,可以生成大量的 3 字词和少量的 4 字词. 在汉语中,具有这类功能的字是很有限的,可以通过 Suffix 词典对尾字的识别过滤垃圾串.

① 对于切分好的语料库 C,寻找 2+1,3+1 模式的串,并统计次数.

② 计算每一个字在该模式串中作为尾字出现的次数.

③ 统计出次数最高的 N 个字,作为我们的尾字集合 $S(T)$.

④ 对于 2+1,3+1 模式的候选新词,如果尾字出现在 $S(T)$中,那么该串是新词;否则转到下一个候选新词.

3.3.3 词性过滤机制

经过单字串垃圾过滤和尾字过滤,候选新词中还会剩余部分命名实体、数量词、时间词和介词短语等垃圾,我们通过使用词性过滤规则进行过滤. 词性过滤的规则我们是自动从语料库中学习得来的,只要有足够多的切分好的语料,系统就可以不断扩充自己的规则库.

3.3.4 使用独立词概率过滤 1+1 模式垃圾串

独立词概率是度量一个字在句子中独立成词的可能性的指标. 给定语料 C,对于字 c,$N(c)$表示 c 出现的次数,$N(word(c))$表示 c 独立成词的次数,$IWP(c)$表示 c 的独立词概率,那么:

$$IWP(c) = \frac{N(word(c))}{N(c)}, \quad (1)$$

$IWP(w) = IWP(c1)\,IWP(c2)$,其中词 $w = c1c2$.

如果 w 的独立词概率越大,w 是垃圾串的概率越大. 经实验可知,独立词概率对于二元词的识别有较好的效果,但对于多元词的识别效果不好,所以我们使用独立词概率进一步对 1+1 模式的垃圾串进行过滤. 随着准确率的提高,有些新词被过滤掉了,所以召回率有所下降. 我们根据阈值的浮动来调整召回率和准确率.

3.3.5 其他模式串的识别

对于没有明显特征的1+2模式、1+3 模式、1+2+1 模式和 2+2 模式的候选新词,我们设置一个较高的频度阈值,只有出现频度大于该阈值的模式串才作为候选新词处理. 这样也可以对这些模式的串进行部分召回. 对于更长的词串,检测的复杂度增加,我们在本文中不做讨论.

4 性能分析

4.1 准确率和召回率计算公式

$$Precision = \frac{N(Correct)}{N(Detected)} \times 100\%, \quad (2)$$

其中,$N(Correct)$表示正确检测的新词数,$N(Detected)$表示检测到的新词总数.

在大规模语料库上计算新词识别的召回率是比较困难的,因为要人工在该语料库中统计所有的新词代价非常昂贵. 根据我们的限定,在大规模的语料库上,只有重复出现过的串才可能是候选新词. 所以我们设计了适用的新词召回率公式:

$$Recall = \frac{N(Correct)}{N(Correct) + N(Filtered)} \times 100\%, \quad (3)$$

其中,$N(Correct)$表示正确检测的新词数,$N(Filtered)$表示被错误过滤掉的新词数.

4.2 实验结果与对比

我们主要分析垃圾串过滤机制的性能. 首先我们通过语料训练出 4 个词典,然后逐次添加每个词典进行性能对比,结果如表 1 所示.

语料库规模:400 张网页.

Table 1 Comparison of Four Lexicons

表 1 使用 4 个过滤词典的性能比较

Parameter	Garbage	+ Garbage Head	+ Garbage Tail	+ Suffix
New Word	88	72	60	56
Correct	50	48	47	45
Wrong	5	7	8	10
Recall (%)	90.9	87.3	85.4	81.8
Precision (%)	56.8	66.7	78.3	80.4

Note: The results are based on the POS Filtering.

语料库的内容和规模都会影响到新词检测的准确率和召回率,比如,对我们的系统而言,语料库越

大,准确率和召回率越大.不同的方法对新词的定义不同,也会影响到新词检测的结果.所以我们提出的新词检测方法和前人的工作缺乏一个统一的平台进行对比,于是设计了一个基准实验来评价垃圾串过滤机制的整体性能.在基准实验中,新词检测的依据是出现频度,而不进行垃圾串的过滤.结果如表 2 所示.

语料库规模:650 张网页.

Table 2 Comparison with the Baseline
表 2 基准实验和本系统的对比

Parameter	Baseline	Our System
Document Frequency	1	1
Total Frequency	2	2
Garbage Filtering	no	yes
New Word	602	202
Correct	192	162
Wrong	0	30
Recall (%)	100	84.31
Precision (%)	31.9	80.24
F-Measure (%)	48.4	82.22

我们在 ICTCLAS 自动切分的基础上,对这些网页中的新词进行了人工标注.为了计算新词识别的准确率和召回率,我们把被过滤掉的重复串进行存储和统计.最后发现在 202 个新词中,包含 1+1 模式的新词 77 个,1+1+1 模式的新词 13 个,1+1+1+1 模式的新词 3 个,2+1 模式的新词 52 个,3+1 模式的新词 4 个,英文词 13 个,共计 162 个.在错误过滤掉的新词中,2+1 和 3+1 模式的词较多.因为我们的词缀表只包括那些频繁作为词缀的字,有些新词的词缀并没有包含在词缀表中,导致未被检测到,通过扩大词缀表的规模,系统的召回率还会有所上升.对于单字串模式的新词,由于垃圾串过滤机制非常有效,准确率和召回率都比较高.另外,为了便于统计系统的召回率,在该实验中使用的语料库较小,如果加大语料库规模,准确率和召回率都会有所提高.

在大规模实验中,我们采集了 70 万张网页,大约 17GB 的语料,然后设置文档频率阈值为 50,总频率阈值为 100,最后检测到新词 568 个.经过人工筛选得出,准确率达到 95% 以上.为了减少人对新词检测过程的干预,我们更关心的是系统的准确率,而系统召回率的小幅度下降,可以通过大规模的语料库来弥补.

列举我们找到的部分新词:

流调 讨薪 减仓 超跌 普跌 重仓
婚检 独派 诉求 反制 桥吊 世卫
宪改 东突 国亲 海归 军演 彩民
双规 军购 考录 反恐 非典 泡吧
善后款 采购团 通行费 彩票业
疾管局 偷私渡 男人婆 妖魔化
点击率
统独公投 反独促统 西电东送
非典疑似 阴胜阳衰 央视春晚

5 结论和下一步工作

本文提出的方法在因特网上的大规模生语料库基础上,使用重复串搜索技术、垃圾串过滤技术和独立词概率技术,进行在线新词检测,并取得了较好的效果.新词检测的全过程,包括网页采集、垃圾词典的生成、过滤规则的生成和独立词概率的获取都是自动完成的,相对于人工提取规则的系统而言,大大减少了人的工作量,系统在新词检测、术语库建立、热点命名实体统计等领域具有较高的应用价值.

下一步对新词检测及相关工作的研究,有以下几个重点.①1+2,1+3,1+2+1 和 2+2 等模式的新词,虽然数量较少,但仍需要寻找更有效的检测方法.②提高低频新词的检测性能.③如果我们能够准确地获取语料的出现时间,而且时间分布比较均匀连续,我们可以考虑引入时间参数,作为新词检测的一个重要手段.④在保证较高的新词识别准确率基础上,对新词相关信息进行提取,发掘新词和新词、新词和旧词之间的语法和语义相关性,分析出与新词相关的社会现象、新词的语境,以指导对新词的理解和使用.

参 考 文 献

1 K. J. Chen, Ming-Hong Bai. Unknown word detection for Chinese by a corpus-based learning method. International Journal of Computational Linguistics and Chinese Language Processing, 1998, 3 (1): 27~44

2 K. J. Chen, W. Y. Ma. Unknown word extraction for Chinese documents. The 19th COLING 2002, Taipei, 2002

3 Jianfeng Gao, Mu Li, Andi Wu, *et al*. Chinese word segmentation: A pragmatic approach. Microsoft Research, Technical Report: MSR-TR-2004-123, 2004

4 Nie Jian-Yun, Wanying Jin, Mareie-Louise Hannan. A hybrid approach to unknown word detection and segmentation of Chinese.

Int'l Conf. Chinese Computing, Singapore, 1994
5 Hua-Ping Zhang, Qun Liu, Hao Zhang, *et al*. Automatic recognition of Chinese unknown words based on roles tagging. The 1st SIGHAN Workshop on Chinese Language Processing, Taipei, 2002
6 Zheng Jiaheng, Li Wenhua. Internet new words according to word-building rule. Journal of Shanxi University(Natural Science Edition), 2002, 25(2):115~119 (in Chinese)
(郑家恒, 李文花. 基于构词法的网络新词自动识别初探. 山西大学学报(自然科学版), 2002, 25(2):115~119)
7 Andi Wu, Zixin Jiang. Statistically-enhanced new word identification in a rule-based Chinese system. The 2nd Chinese Language Processing Workshop, Hong Kong, 2000
8 Fuchun Peng, Fangfang Feng, Andrew McCallum. Chinese segmentation and new word detection using conditional random fields. COLING 2004, Geneva, Switzerland, 2004
9 Min-Jer Lee, Chien-Kang Huang, Lee-Feng Chien. Automatic construction of a bilingual live dictionary for spoken language processing applications. Oriental COCOSDA99, Taipei, 1999
10 Zou Gang, Liu Yang, Liu Qun, *et al*. Internet-oriented Chinese new words detection. Journal of Chinese Information Processing, 2004, 18(6):1~9 (in Chinese)
(邹纲, 刘洋, 刘群, 等. 面向 Internet 的中文新词语检测. 中文信息学报, 2004, 18(6):1~9)

Cui Shiqi, born in 1981. Master candidate. His main research fields: natural language processing and new words detection.
崔世起,1981 年生,硕士研究生,主要研究方向为自然语言处理、新词检测.

Liu Qun, born in 1966. Doctor of science Professor. His main research fields: natural language processing, machine translation and information extraction.
刘群,1966 年生,博士,研究员,主要研究方向为自然语言处理、机器翻译、信息提取.

Meng Yao, born in 1970. Doctor of engineering. Her main research fields: natural language parsing and machine translation.
孟遥,1970 年生,博士,主要研究方向为自然语言句法分析、机器翻译.

Yu Hao, born in 1971. Doctor of engineering. His main research fields: natural language processing and network information processing.
于浩,1971 年生,博士,主要研究方向为自然语言处理、网络信息处理.

Nishino Fumihito, born in 1956. Master of engineering. His main research fields are machine translation and natural language processing.
西野文人,1956 年生,硕士,主要研究方向为机器翻译、自然语言处理.

Research Background

Segmentation is a basic task in Chinese information processing. The detection of new words is a difficult problem to solve in segmentation. This paper gives an effective solution to this problem. In comparison with the previous study, the paper presents a new approach to detecting words based on large-scale corpus. It first collects plenty of Web pages from the Internet, transforms them into plain text, segments them with ICTCLAS, calculates the frequency of the strings, and designs different filtering mechanisms to separate true new words from garbage strings. While filtering, it discovers rich features of various new word patterns, and produces three garbage lexicons and a suffix lexicon automatically through the corpus, and achieves good results in detection. In future research, we will consider more kinds of new word patterns, and try to detect new words with low frequency. To add a time parameter may be a complementary approach to detecting new words. We will also find relations between new words and known words, to make full understanding of the new words. Our research work is supported by the National High-Tech Research and Development Plan of China under Grant No. 2004AA114010, No. 2003AA111010, and Fujitsu Research & Development Center.