基于WWW的未登录词识别研究

WWW-based Recognition of Non-login Words

韩 洁 周 勇 刘少辉 史忠植
(中国科技大学研究生院 北京100039)

Abstract Currently, very little reference material can be found on the research of non-login word recognition. Solutions based on rules and syntaxes can't satisfactorily solve all kinds of problems of non-login word recognition. This paper will study and compare several existing solutions. The proposed solution is to extract N-grams after words separation, from which non-login words can be extracted by means of probability statistics. Experiments have demonstrated that this method has favorable efficiency, recall ratio, and accuracy.

Keywords Non-login word, Recognition, N-gram, WWW

一、引言

当前,随着国民经济信息化的不断发展以及Internet的普及应用,全世界丰富的信息资源展现在我们每个人面前。如何从大量的信息中迅速有效地提取出所需信息极大地影响着我国计算机技术和信息技术的发展和应用推广。据统计,在信息领域中,80%以上的信息是以语言文字为载体的,因此,中文信息处理技术成为我国重要的计算机应用技术。未登录词的识别是中文信息处理技术中的难点之一。它在Internet数据挖掘、信息检索、图书馆图书文献管理、语音识别等应用中有着非常重要的作用。

未登录词是指中文分词处理中没有包含在分词词典中的词。它可分成以下几类:

1 人名:如:张建国
2 地名:如:洛阳市
3 组织机构名:如:派出所、复旦大学、某某公司
4 音译词:如:索尼、英特尔
5 缩略语:如:足协
6 专业术语:如:域名
7 日常用语:如:打的
8 新产生的词汇:如:疯牛病、DIY

改革开放以来,中国获得了前所未有的变化和发展,汉语的词汇系统也处于推陈出新的动荡期[1]。这些新词语活跃在信息敏感的Internet上。它们对词频统计和文档类别的识别有重要影响。有些系统放弃处理未登录词,这是不可取的。例如:一篇体育新闻,内容是介绍国际大赛中普遍服用违禁药品现象的文章,其中未收录到分词词典中的词"禁药"出现9次,而其它和体育相关的词最高出现3次,因此,如果不能正确地识别出"禁药"一词并统计其出现频率,使用基于VSM模型的分类器时就会对这篇文章的分类产生较大影响。已知词汇和未登录词在向量模型的构成、文档分类和检索中占有同等重要的地位,未登录词的识别成为文档处理中的关键问题。

二、常用的处理方法

1. *基于规则和知识库的方法* 首先收集大量的某类未登录词,建立用词表,如《中国地名录》等,然后根据各类识别规则分别对其进行判断[3]。例如同济大学的应志伟、柴佩琪等人开发研制的文语转换系统就是利用这种方法识别人名的。这种方法查全率和查准率非常高,但也存在一些缺陷:识别效果的好坏极大地依赖所利用的资源是否全面、科学、有权威性。如果占有的资源比较少,覆盖面比较小,势必会影响识别的效果[1]。因此它需要十几万甚至几十万的词条,以保证统计结果的全面可靠。例如《中文文本自动识别与分类》一书以中国地名委员会编纂的《中华人民共和国地名录》为基础建立中国地名库和《中国姓名录》中收集的十几万人名统计其中的用字规律。第二,这种方法可识别的词的类型仅限于具有统计特征的几类词,而对于大量的普通词汇和新词汇却没有作用。

2. *基于概率统计的方法* 语料库语言学的发展促使计算机语言学家们越来越重视数理统计在语言学中的应用。基本思路是:对语料库中相邻的各个字的组合的频度进行统计,计算它们的互现信息。互现信息体现了汉字之间结合的紧密程度。这种方法只需对语料库中的字的组合频度进行统计,不需要依赖词典,因此它在一定程度上能识别未登录词;但也有一些局限性——即对常用词的识别精度差,时空开销较大。微软中国研究院的Zhang Jian、Gao Jianfen、Zhou Ming等人在《An Experimental Study on a Very Large Corpus》论文中提到的"mutual information and context dependency"方法就是对这种方法的一种改进,据该论文实验数据表明,效果比较好。

三、N-grams基本概念

N-grams方法被广泛地应用于中文等亚洲语种的信息检索(INFORMATION RETRIEVAL)系统中,它的优点是无需语法知识。其基本原理如下:把一组字符串按固定长度切成一个一个的单元,最常使用的是一元组(uni-grams)、二元组(bi-grams)和三元组(tri-grams)。例如:字串ABCDE(每个英文字母代表一个汉字),其一元组形式为:A/B/C/D/E,二元组形式为:AB/BC/CD/DE,三元组形式为:ABC/BCD/CDE。在中文IR系统中,常以N-grams为索引项进行信息检索,从而避开由汉语分词的复杂性和不准确性对信息检索带来的影响。

N-grams可从原始文本(未经过分词)中抽取,也可从分词后的文本中抽取。我们采用刘开瑛在《中文文本自动识别与分类》一书中对N-grams的定义以及抽取方法。

例 (摘自北京晚报)

原文:来自河南驻马店的民工刘新兵参加了铁路价格听

韩 洁 硕士研究生,主要研究方向为信息检索、文档聚类。周 勇 硕士研究生,主要研究方向为自然语言理解。刘少辉 博士研究生,主要研究方向为数据挖掘、信息检索。史忠植 研究员,博士生导师,主要研究方向为人工智能、知识工程。

证会。

初步分词:来自 河南 驻 马 店 的 民 工 刘 新 兵 参加 了 铁路 价格 听证 会。(人名和地名等没有识别出来)

正确分词:来自 河南 驻马店 的 民工 刘新兵 参加 了 铁路 价格 听证会。

我们从初步分词的文本中提取 n 元组,结果为:

2元组:驻马 马店 店的 的民 民工 工刘 刘新 新兵 听证 证会(10个)

3元组:河南驻 驻马店 马店的 店的民 的民工 民工刘 工刘新 刘新兵 兵参加 参加了 了铁路 价格听 听证会(13个)

可以看到,我们希望切分出来的人名"刘新兵"、地名"驻马店"以及常用词"民工""听证会"都包含在 n 元组中,剩下的就是如何将这些词提取出来。

四、从 N-grams 中抽取未登录词

1 剔除噪音字 对汉字词的统计表明,不同汉字同其他字结合成词的能力是不同的。一个字同其他字结合成词的次数越多,该字的构词能力就越强。反之,则为构词能力低的字。但构词能力低的字出现的频率就低。例如:指示代词"我"、"你"、"它",助词:"的"、"了"等在文档中出现的频率非常高,并且和其它字构成常用字组,如:"这是"、"我的"…。我们称这些出现频率高但构词能力低的字为噪音字,在统计 N-grams 时剔除这些噪音字是非常必要的。在第三节的举例中,我们假定"的"和"了"为噪音字,包含噪音字的 N 元组被去掉,则剩余的 N 元组是:

2元组:驻马 马店 民工 工刘 刘新 新兵 听证 证会(8个)

3元组:河南驻 驻马店 民工刘 工刘新 刘新兵 兵参加 价格听 听证会(8个)

注:(带下画线的词是我们希望抽取出来的未登录词)

2 重叠处理 有两种类型的重叠:第一:三元组"刘新兵"包含了2元组"刘新"和"新兵";第二:在三元组中,"刘新兵"和"兵参加"两个元组有重叠字出现。为消除这种因素,使用如下过程来剔除:比较重叠元组的出现频率,通常保留出现频率更高的为候选词。而对于包含情况,假设 n 元组 X 包含在更长的元组 Y 中,则用公式 freq(X)=freq(X)−freq(Y)重新计算元组 X 的频率。在实践中这种方法很有效。

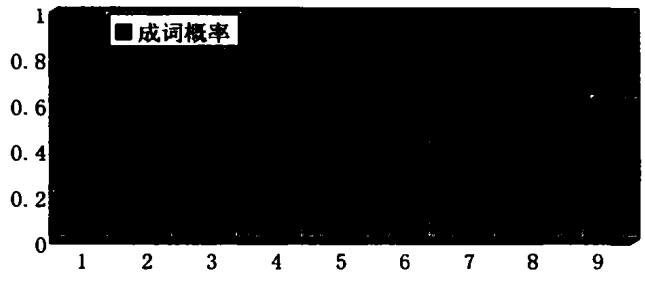

图1 出现频率 f 和成词概率(1000篇文档2954 Kbyte)

3 确定未登录词 用 N-grams 的出现频率来判断是否成词:通过统计,我们发现,如果某个 N-grams 多次出现在某文档中,那么它可能是一个相对固定的词。通过对28类1000篇文档(2954Kbyte)的统计,我们发现,对于平均长度为3-4kbyte 的网页,当 n 元组词频>=f(f=7)时,成词的概率即可达到我们满意的程度(85%)。用出现频率来判断是否成词的方法在处理大规模文档时,算法简便,速度很快,而且准确率也令人满意(图1:横坐标表示出现频率,纵坐标表示成词概率)。

表1 筛选出来的未登录词举例

出现频率>=7的未登录词	由互信息公式统计出来的词
夏利(出现频率=7)	搜狐(出现频率=1)
降解(出现频率=15)	免费邮箱(出现频率=2)
王师傅(出现频率=7)	毛宁(出现频率=3)
罗布泊(出现频率=16)	镜像(出现频率=2)
市场经济(出现频率=7)	财经(出现频率=4)
纳斯达克(出现频率=8)	证券(出现频率=2)

对于出现频率小于域值 f 的词,可用互信息(mulual information)公式来筛选:互信息是信息论中的概念,通过对大规模数据的统计信息来描述两个字符串之间的相互依赖关系。在汉语分词中常用它来描述字与字之间的紧密程度,当紧密程度大于某域值 t 时,我们就判定字与字之间可以组合为词(见表1)。

表2 查全率(%)和查准率(%)

	查全率(Recall%)	查准率(Precision%)
出现频率>=7	90.9%	85.9%
互信息(MI)	71.6%	74.02%
总体	80.01%	79.19%

注:关于以上三个表格中的数据的三点说明:

1 以上三个表格中的数据和切分词典的大小有关,我们使用的切分词典约有十万词条;

2 汉语分词规范还有待统一,一个词组是否成为独立的词,不同专家对它的看法不一致;它影响我们对查全率和查准率的计算;

3 文章的长度越短,得到的结果的准确率越低。

结论以及未来的工作 当前,关于未登录词的识别的研究资料可参考的资料非常少,在基于规则和语法知识的方法不能很好地解决未登录词识别的问题的情况下,本文研究比较了现有的各种方法,在分词之后提取 N 元组,通过概率统计手段从中提取未登录词。实验表明该方法的效率、查全率和查准确率都是令人满意的。算法中各个域值的设定、中文词组的确定规则以及噪音字的选取对该方法的精度影响很大,值得我们做进一步的研究。公正地说,统计方法并不是包治百病的良药,而是人类由于认识能力和认识范围的有限迫不得已采用的方法(而且是唯一的方法)[4]。

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

2 刘少辉,董明楷,张海俊,李蓉,史忠植. 一种基于向量空间模型的多层次文本分类方法. (该论文已被中文信息学报录用)中科院计算所智能信息处理开发实验室

3 黄萱菁. 大规模中文文本的检索、分类于摘要研究:[复旦大学博士学位论文]. 1998

4 于江生. 计算语言学中的概率统计方法. 北京大学计算语言学研究所,1999,9

5 Zhang Jian,Gao Jianfen,Zhou Ming. An Experimental Study on a Very Large Corpus. Microsoft Research,China

6 李国臣. 文本分类中基于对数似然比测试的特征词选择方法. 中文信息学报,1999,13(4):16~21

7 Gotoh Y,Renals S. Variable Word Rate N-GRAMS. University of Sheffield,Department of Computer Science

8 姚天顺,等. 基于规则的汉语自动分词系统. 中文信息学报,1990,1