基于“动态流通语料库”进行“有效字符串”[1]提取的初步研究

隋岩　张普
北京语言文化大学语言信息处理研究所 北京 100083
E-mail:zhangpu@blcu.edu.cn
suiyan@blcu.edu.cn

摘要：本文提出了“有效字符串”的概念，试图找到从大规模中提取这种字符串的新方法。主要是以“流通度”理论为核心，通过对语料进行周遍切分并辅以“复合词典系统”，计算字符串的“流通度”。最终得到一个能够动态更新的词表。目前的方法还是试验性的，并且借鉴了前人丰富优秀的研究成果。
关键词：流通度、字符串、语料库

A Preparatory Study on Distilling “Valid Character Strings” Based on “Dynamic Corpus”

SuiYan　ZhangPu
Language Information Processing Institution, Beijing Language and Culture University, Beijing 100083
E-mail: zhangpu@blcu.edu.cn
suiyan@blcu.edu.cn

Abstract: This paper suggests a new concept of “valid character string” and gives a new method of distilling valid character strings in terms of circulation. We try to gain a dynamic updated glossary. Now we develop a testing method by referencing previously work of other people.
Keyword: CIRCULATION，CHARACTER STRINGS，DYNAMIC CORPUS

汉语信息处理的一个最基本然而也是最让人头疼的问题就是自动分词。近 20 年来人们为此付出了不懈的努力，从不同的角度采用多种方法进行探索。例如众所周知的规则法、统计法以至于目前比较流行的规则加统计法等等。这些方法各具特色，但是都没有从根本上解决问题。“计算机……要像人一样对句子进行处理，就必需把这一串字符切成合乎人的语感的一

[1] 本文承教育部科学技术研究重点项目资助。

串词。这几乎是我们进行其它所有跟自然语言处理相关的应用开发，诸如机器翻译、人机对话等的前提。”“但在实践中仍有相当多的分词歧义问题、未定词问题等困扰着研究人员”（詹卫东，2000）。本文亦无意正面接触这个难题，而是试图另辟蹊径，探索一下有别于传统思路的新方法。

1. “有效字符串”与“高频字符串”

汉语信息处理研究中“词”的问题，从本质上讲就是字符串的问题。任何一个系统，如果能够解决这个问题，也就成功了多半，这是不言自明的。把词当成字符串来看待的观点，越来越为人们所接受。一方面是因为汉语的词、词组、短语甚至句子之间的界限不是很明晰，另一方面，大量新词新语新用法的出现，也带来了不小的混乱，使得词语界限的划分和确定变得越来越困难。针对“高频字符串”的概念，我们提出一个与其相对应的概念“有效字符串”来解决从大规模真实文本中提取动态词表的问题

采用 N 元语法模型的方法可以从大规模真实文本中统计出“高频字符串”，但是高频的未必是有效的。所谓“有效字符串”一定有相对完整的意义并且具有一定水平流通度，这样才可能成为一个相对独立的语法单位（单纯的或者复合的）。我们所要提取的“有效字符串”总体上包括两个部分：规范的和非规范的。在大规模真实文本中占大多数的是那些已经为各类词典收录的词或者词组，然而有一个事实不容忽视，那就是新的词语新的用法几乎每天都在产生。它们中的大多数还不能马上被人们接受，不过随着时间的推移有一些就存在下来了，甚至被作为传统意义上的“词”和“词组”收入词表或者词典之中去。

采用 N-gram 方法对大规模真实文本进行分词处理（只考虑 bigram 模型）时，就会产生大量的“切分垃圾”，即“无效字符串”。这一点已经得到了研究者的充分证明（宋柔等，1997，1998）。这些“垃圾”中，有相当一部分都是“高频”的，然而它们却绝对不是我们想要的字符串。如果把它们汇入到词典或者系统中去，只能起到干扰和“污染”的作用。

2. “动态流通语料库”与“有效字符串” 提取

尽管我们的研究目标不是自动分词，但是跟自动分词也不是没有关系。处理词汇或者字符串，前提是从大量的语料中把它们提取出来。

我们的方法就是通过对人们语感的模拟，对大规模真实文本进行历时的考察和监控，排除字符串切分过程中产生的大量“垃圾”，找到“有效字符串”。

语感本身只能模拟而不能计算，因为语感是不可以直接量化的，而语感的模拟却通过计算来实现。对于计算机来说，真实文本无论规模大小，都是没有任何意义的。构成文本的字、

词、句都是符号或符号串，我们无法让计算机去“理解”什么（当然，这有赖于我们如何定义“理解”）。唯一可以做到的就是让它按照某种算法（步骤）处理字符。我们通过计算方法让计算机模拟了语感，当然不是意味着机器已经获得了像人一样的语感。不过我们可以利用跟语感相关的机制对计算的结果进行评估，尽量接近和模拟人脑处理语言的心理过程。

张普（1999）正式提出“流通度”这个概念到现在，将近两年的时间已经过去了。我们的研究工作一直没有停顿。正在着手创建用于“启动部分”动态流通语料库。

从构成上讲，“动态流通语料库”可以分为两大块：“累计部分”和“滚动部分”。这也体现语言现象的两个属性，共时性和历时性。“累计部分”是历时的，语料库中所有的语料都有时间属性和以时间点为刻度的流通度属性，我们称之为“语料仓库”；“滚动部分”是共时的，我们称之为“滚动语料库”。它以一定的时段为更新周期，截取真实文本的一个“横断面”，以“流通度”为手段加以处理，完成之后即滚动进入“语料仓库”之中。在这样的处理机制下，“滚动部分”的语料规模可以很大，也可以不大，关键是我们所确定的体现滚动周期的时段颗粒度。

“语料仓库”和“滚动语料库”在结构上是同构的、对等的。这样，滚动部分处理过的语料才能顺利进入语料仓库形成具有历时意义的“动态流通语料库”的主体，才能够通过流通度的计算对字符串进行评估，从而找到我们需要的“有效字符串”。

新词的发现和提取是语言信息处理研究的一个重要组成部分。在中文信息处理中，这个问题尤为突出。因为中文句子的书写方式是词与词连写，词与词之间没有明确的界限标记（例如空格、符号等），将词典中已有的词从句子中准确地切分出来已经是一件很不容易的事情，更不用说找出那些新出现的和词典中未登录的词了。有学者建议实行信息处理用的汉语“分词连写”方式，人们的意见还不能统一，其结果如何也不得而知。

前文论及，中文信息处理中自动分词的研究时间已经很长了，研究者们进行了多方位的尝试，寻找自动分词的规律、提出各种各样的算法。然而时至今日，问题仍未得到令人满意的解决，距离取得突破的时刻仍然遥遥无期。

本研究的焦点亦不在自动分词上，也不想奢望取得分词研究的突破。而是关注真实语料中有“效字符串”的发现和提取。依据“流通度”的理论和方法，从一个全新的角度处理开放的真实语料，不仅要发现和提取字符串（包括新的普通词汇和新的专名），而且要对这些提取出来的字符串进行流通度评估，按照流通程度编排出能够动态更新的词表。

3. “动态流通语料库”方法

“有效字符串”的发现和提取无法超越对语料切分这一必不可少的步骤。但是，由于目标不同，自动分词中的很多棘手问题（如“歧义切分”等）本研究并不会遇到。本研究所强

调的主要是找到中文真实语料中具有完整意义的字符串并用流通度原则对它们进行评估，以便确定这些字符串能否作为进入词典。这种统计比单纯意义上的“高频字符串”提取要全面得多。

有了动态流通语料库，配合“复合词典系统”，我们有望从大规模真实文本中高效准确的获取那些对中文信息处理用处极大的“有效字符串”。

“复合词典系统”是不同类型的底表结合在一起的产物。这些底表之间存在着连接关系，共同协作完成“有效字符串”的评估和提取。

底表（词典）在分词系统中的地位是举足轻重的。底表越大，分词准确性原则上讲也就越高。在本项研究中，我们准备采用较大规模的复合词典系统，对切分结果进行综合处理。

复合词典系统包含了一系列相关联的底表，目前考虑的有以下五种：

1)基本底表词典。(2 字、3 字……8 字)

2)“垃圾字串”词典

3)最小词类实例词典（有穷词类）

4)单音节字符词典（非成词字符）

5)专名（人名、地名、机构名等）词典

关于“垃圾字串”词典我们有这样的认识：过去人们总以为切分过程中产生的所谓“垃圾串”出了影响系统的正常运转之外根本就不存在什么积极的作用。我们认为不然，垃圾字符串并不是完全无用的东西。原因很简单：如果系统知道了什么是无用的，必然就能够分辨出什么是有用的。更为重要的是，“垃圾字符串”也是有流通度的。本研究目前已经拥有一个从 2 亿字语料中筛出的“垃圾字串（两字）”表作为“垃圾字串”词典的基本构成部分；并且已经补充了从 100 万字中筛出的 3 字“垃圾字串”。在以后的研究中，这个词典还会逐步扩大。开始的时候垃圾字符串大量出现，但到了一定的数量之后，增长的幅度就会越来越小。

上述这些词典还不是整个词典系统的全部，既可能扩大范围，又能够动态更新。所有的词典都能从动态流通语料库中获得新的内容，因而它们是与动态流通语料库共同生长的。

4. 文本处理

动态流通语料库是整个处理过程的核心，实现“有效字符串”的发现和提取并不是一个简单的、单一的过程，也不是一次处理就能完成的。大规模真实文本既是动态流通语料创建的基本素材，也是有效字符串发现和提取所依据的知识库。

现在，我们还没有个真正意义上的“动态流通语料库”，要进行新词的发现和提取还有相当的困难。但是，由于进行了比较充分的前期准备工作，我们可以在一定范围内展开小规模的试验。本次试验的真实文本规模如下：

以国内公开出版发行的报纸为文本选取对象，从媒体流通度最高的前300种报纸中选出十种；时间跨度为1999年1月1日到1999年12月31日；每种报纸至少选定5个版面，最大限度地做到体裁和题材的均衡分布，预计总量在1亿词次左右。将这些报纸按季度划分为4个部分，作为流通度评估的时间刻度（滚动周期），对各个部分的数据分别进行处理。

整个处理过程算不上复杂，不过确实是有些繁琐。传统方法因为要进行互信息和概率计算，为了简化计算过程，对于N-元模型往往只考虑到N=2或N=3（Bigram模型和Trigram模型）的情况。在我们的方法中，并不去计算这些复杂概率关系，因而我们能够对语料进行周遍的多次切分（目前主要考虑N=2至N=8。在初步的实验中发现，相对比较复杂的情况应该出现在N=2、N=3和N=5上，主要是N=3），以便得到真实文本中包含的所有可能的字符串。当然，仅仅进行这样的切分是没有什么意义的，因为在切分结果中，无用的垃圾串占了绝大部分(宋柔教授在二字接续研究中有详细的说明)。问题的关键就是怎样从这些浩如烟海的“垃圾”掩盖之下找到真正的“金子”（有效字符串）。

由于时间所限，我们目前只对120多万字的真实文本进行了N=2和N=3的切分。得到了两个（处理前的）切分字符串表：二字串大约109万个，三字串大约90万个，两项合计将近200万个。当然，这只是所要处理的全部语料中极其微小的一部分。

算法：

因为不涉及到概率计算，整个切分过程比较单纯，实际上只是纯粹的数目计算。

1)切分语料（N=2至N=3）

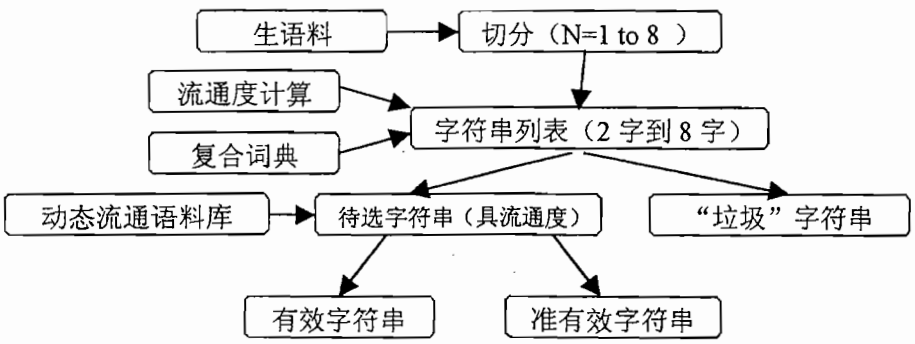
2)生成“可能字符串”列表。该列表应包含如下内容：字符串、字符串重现次数、字符串长度。

3)进入复合词典系统进行处理。滤掉“垃圾字符串”，标记“有效字符串”，标记“非垃圾字符串”和“非有效字符串”，记作“准新词字符串”。

4)依据流通度原则，对这些有效字符串进行评估，重新排序。

5)给出“准新词字符串”列表，为下一个滚动周期的对比作准备。

整个加工提取流程如下：



需要特别强调的是，采用“流通度”和“动态流通语料库”的方法进“有效字符串”提取的一个最突出的特点就是引进了“历时性”这个观念。一个滚动周期的处理可能发现不了某些“有效字符串”，不过，经过几个周期之后，如果某个字符串确实是有效的，就能够被提取出来。

上述流程图中的“生语料”部分，在一个滚动周期之后，就会在字符串一级被标上“流通度”标记，追加到“动态流通语料库”之中。

5. 结语

目前我们正在对大规模语料进行处理，预计 3-4 个月就会有具体数据得出。所有数据会作为本文附录提交给本次会议。

这种方法对语料的切分本质上是属于无词典分词的范畴。不过不同的是，我们并不是仅仅在语料内部寻找提高分词准确率的统计学依据，而是充分运用已有的词典系统和“切分垃圾”从外部进行正向和反向排除操作。尤其应该说明的是，这个处理过程并不是一次性的，它在“动态流通语料库”的每一个滚动周期中都存在，随着语言现实的不断发展而动态地进行。

参考文献

[1]张普，关于语感与流通度的思考，《语言文字应用》，1999.2

[2]张普，关于大规模真实文本语料库的几点理论思考，《语言文字应用》1999.1

[3]韩客松、王永成、陈桂林，无词典高频字串快速提取和统计算法研究，《中文信息学报》，2001.2

[4]宋柔、戴伟长等，现代汉语二字结构工程，《ＩＣＣＩＰ９８国际会议论文集》

[5]吴应良、韦　岗、李海洲，基于字统计语言模型的汉语语音识别研究，《计算机应用研究》，2000.5

[6]孙茂松、卢红娜、邹嘉彦，基于隐 Markov 模型的汉语词类自动标注的实验研究，《清华大学学报》(自然科学版)，2000.9

[7]詹卫东，80 年代以来汉语信息处理研究述评，《当代语言学》，2000. 2

[8]周　强，一个汉语短语自动界定模型，《软件学报》第 7 卷增刊，P315～322，1996

[9]魏　欧、吴　健、孙玉芳，基于统计的汉语词性标注方法的分析与改进，《软件学报》，2000.4

[10]冯志伟，信息处理用汉语分词连写，《语文建设》，2001.3

[11]刘开瑛，中文文本自动分词和标注，商务印书馆，2000

[13]应志伟、柴佩琪、陈其晖，文语转换系统中基于语料的汉语自动分词研究，《计算机应用》，2000.2

[14]姚天顺等，基于规则的汉语自动分词系统，《中文信息学报》1989.4