ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.3, March 2006, pp.356−363
DOI: 10.1360/jos170356

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向信息检索的自适应中文分词系统*

曹勇刚+, 曹羽中, 金茂忠, 刘 超

(北京航空航天大学 计算机学院,北京 100083)

Information Retrieval Oriented Adaptive Chinese Word Segmentation System

CAO Yong-Gang+, CAO Yu-Zhong, JIN Mao-Zhong, LIU Chao

(School of Computer Science and Engineering, BeiHang University, Beijing 100083, China)

+ Corresponding author: Phn: +86-10-82324488 ext 885, E-mail: ygcao@cse.buaa.edu.cn, http://sei.buaa.edu.cn

Cao YG, Cao YZ, Jin MZ, Liu C. Information retrieval oriented adaptive Chinese word segmentation system. *Journal of Software*, 2006,17(3):356−363. http://www.jos.org.cn/1000-9825/17/356.htm

Abstract: New words recognition and ambiguity resolving have vital effect on information retrieval precision. This paper presents a statistical model based algorithm for adaptive Chinese word segmentation. Then, a new word segmentation system called BUAASEISEG is designed and implemented using this algorithm. BUAASEISEG can recognize new words in various domains and do disambiguation and segment words with arbitrary length. It uses an iterative bigram method to do word segmentation. Through online statistical analysis on target article and using the offline words frequencies dictionary or the inverted index of the search engine, the candidate words selection and disambiguation are done. On the basis of the statistical methods, post-process using stopwords list, quantity suffix words list and surname list are used for further precision improvement. The comparative evaluation with the famous Chinese word segmentation system ICTCLAS, using news and papers as testing text, shows that BUAASEISEG outperforms ICTCLAS in new words recognition and disambiguation.

Key words: word segmentation system; word segmentation algorithm; information retrieval; new word recognition; disambiguation

摘 要: 新词的识别和歧义的消解是影响信息检索系统准确度的重要因素.提出了一种基于统计模型的、面向信息检索的自适应中文分词算法.基于此算法,设计和实现了一个全新的分词系统 BUAASEISEG.它能够识别任意领域的各类新词,也能进行歧义消解和切分任意合理长度的词.它采用迭代式二元切分方法,对目标文档进行在线词频统计,使用离线词频词典或搜索引擎的倒排索引,筛选候选词并进行歧义消解.在统计模型的基础上,采用姓氏列表、量词表以及停词列表进行后处理,进一步提高了准确度.通过与著名的 ICTCLAS 分词系统针对新闻和论文进行对比评测,表明 BUAASEISEG 在新词识别和歧义消解方面有明显的优势.

关键词: 分词系统;分词算法;信息检索;新词识别;歧义消解
中图法分类号: TP391 文献标识码: A

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2004AA112030 (国家高技术研究发展计划(863))

Received 2005-08-02; Accepted 2005-10-11

从句子中划分出的每个有独立意义的词被称作分词.由于中文词与词之间没有明确的边界,因此,中文分词是机器翻译、分类、主题词提取以及信息检索的重要基础.与通用的分词系统不同,面向信息检索的中文分词有着强烈的目的性,对它的评价不应依据人的主观看法,而应该考察其是否有助于提高信息检索的准确度.影响检索准确度的分词结果主要表现在两个方面[1]:对新词的识别(包括人名、地名、组织名和其他不在词典中的术语、俚语或网络用语的识别);歧义的解决(包括交叉歧义和组合歧义的解决).

若对新词的识别能力不够,则会把一个新词拆分为与新词意义不符的词的组合,导致检索新词时,会得到大量无关的仅匹配新词的各个片段的结果.如对“前沿培训网报名须知”,ICTCLAS 分词系统[2]会把它切分为“前沿 培训 网 报名 须知”,这样分词在语法上并没有错误,但搜索引擎因此将会得到大量无关结果(以任意顺序包含这 5 个词的页面都会被找出来).而从语义上来看,前述短语是两个词“前沿培训网”和“报名须知”的组合,用户查找的只是一个特定网站的报名须知.用前述短语对 Google 和百度进行测试(2005 年 7 月 18 日测试),结果均不令人满意.Google 返回了 4 060 个页面,其中只有两个符合要求.而百度只返回了 1 个整句匹配的页面(若手工对查询串分词,百度则返回了 43 篇,同样包含大量不符合要求的页面).可见,在新词识别不佳的情况下,搜索引擎只会走向两个极端:过多匹配或过少匹配.歧义问题可分为交叉歧义和组合歧义,一直是自然语言处理领域的难题.歧义问题相对于新词来说,由于量少,因此对搜索结果影响相对较少,但仍然是一个明显可见的问题.不在词典中的词造成的组合歧义可以归并到新词识别中去.在交叉歧义方面,一些常被提及的词语歧义问题在经过多年的改进后,各主要搜索引擎已解决得几近完美.但百度仍存在明显问题,如用“和服”搜索百度,仍在搜索结果首页出现了“青岛东和服装设备”为标题的页面.可见,包含新词的词语歧义仍然是个潜在的问题.

综上所述,面向搜索引擎的分词系统对新词的识别能力以及歧义的解决都有很高的要求.而经过分析和实验,我们发现目前已有的分词系统在这些方面还是有待改进的.本文提出一个全新的分词系统 BUAASEISEG,它的主要目标是尽可能地解决以上两个问题.通过词典和统计的结合,BUAASEISEG 能够在线进行上下文相关的新词识别和歧义消解.只要具备一定的上下文,它就具有识别各种类型的新词的能力(不局限于人名、地名、组织名)和消解各类歧义的能力.BUAASEISEG 虽然进行了多遍扫描,但它不仅在准确度上得到大幅度提高,而且在速度上也仍然表现良好.

本文的贡献在于:提出了一种全新的基于语境的自适应分词算法,擅于识别长的未知词和消岐;实现了采用本算法的分词系统 BUAASEISEG;结合程序和人工对比,评估了 ICTCLAS 和 BUAASEISEG 的准确度.

下面首先介绍统计模型,然后在系统实现部分介绍 BUAASEISEG 的自适应分词算法,包括分词流程、词频词典和停词、量词表的准备、词频和转移概率的统计和利用.其中给出新词识别和歧义消解的示例;最后给出与 ICTCLAS 分词系统的对比评估以及总结和展望.

1 统计模型

对于一篇文章,忽略其他符号,我们可以把它看作有间隔的字的序列的集合.这里的字,包括中文的字符、中文数字串以及英文的单词和数字串.BUAASEISEG 划分序列的标准是:中文序列被各类非中文字或符号分割,英文序列被非单个空格的其他符号分割.按照以上定义,我们不妨把这样划分出来的序列称为句子.若不考虑句子之间的相关性,则文章 A 是由 m 个序列 $S_i(0<i<=m)$构成的集合:$A=\{S_1,S_2,S_3,\ldots,S_m\}$.每个序列 S_i 由 n 个字 $W_{ij}(0<j\le n)$按序构成,即 $S_i=\langle W_{i1},W_{i2},\ldots,W_{in}\rangle$.对 S_i 可以有多种切分.为了简化表达并保持和其他文献的一致性,我们集中表示 S_i 中的第 j 种切分,称为 $w_{j,1}^{j,m_j}$.其中,m_j 为第 j 种切分中词的个数,第 k 个词为 $w_{j,k}(0<k\le m_j)$.我们还将 W_{i1} 至 W_{in} 表示为 C_1 到 C_n,即 C_1^n.

基于统计的分词算法就是要求下面概率公式的最优解 $\widehat{W}$ [3]:

$$\widehat{W}=\arg\max_{W_j}P(W_j=w_{j,1}^{j,m_j}\mid C_1^n) \tag{1}$$

现在的问题就在于如何选取候选词来进行比较.最一般的做法是按序任取 n 元(*ngram*),然后做所有基于 *ngram* 的全切分比较.由于直接按元划分,可划分出来的词的数目是呈几何级数增长的.如对一个长为 n 的串的

切分进行二元切分可切分出 $n-1$ 个二元组,因此目前最多能做到四元.此时若不考虑任何约束,四元切分的组合数为 $P_1^n \times P_1^{n-1} \times P_1^{n-2} \times P_1^{n-3} = n!(n-1)!(n-2)!(n-3)!$,它是一个非常庞大的排列组合数.当然,由于字串的有序性以及 n 元切分之间的包容和互斥性,实际有效组合数目要少得多.但是,如何利用这些性质来减少排列组合数,以达到可以容忍的程度,并且让这种切分有意义,仍然是值得探索的问题(在系统实现部分,本文提出了全新的基于二元迭代的切分方法).

假设我们选定了切分方法,从而得到了一系列切分,问题就变为如何在这些切分中找到最佳的切分.若把 C_1^n 简称为 S,则有

$$\hat{W} = \arg\max_{W_j} P(W_j \mid S) = \arg\max_{W_j} P(W_j)P(S \mid W_j) \tag{2}$$

由于有了词串,句子就唯一确定了,故 $P(S|W_j)=1$.实际上,我们需要求解式(3)的最优解:

$$\hat{W} = \arg\max_{W_j} P(W_j) \tag{3}$$

如果假设词与词之间独立,可得式(4):

$$\hat{W} = \arg\max_{W_j} P(W_j) = \arg\max \prod_{i=1}^{m_j} P(w_{j,i}) = \arg\max_{W_j} \sum_{i=1}^{m_j} \ln P(w_{j,i}) \tag{4}$$

ICTCLAS 的 n 最短路径粗分概率模型[4]是根据式(4)采用训练集中的全局词频进行求解的,如式(5)所示:

$$\hat{W} = \arg\max_{W_j} \sum_{i=1}^{m_j} \ln P(w_{j,i}) \approx \arg\max_{W_j} \sum_{i=1}^{m_j} \ln(k_{w_{j,i}} / \sum_{l=1}^{m_j} k_{w_{j,l}}) \tag{5}$$

其中,$k_{w_{j,i}}$ 表示词 $w_{j,i}$ 在训练集中出现的次数.但实际情况是,词是和语境相关的,ICTCLAS 只是做到了针对训练集的全局优化,而没有考虑上下文的影响.因此,我们需要引入局部概率的思想,当使用局部概率时,令 $k_{w_{j,i}}$ 为 $w_{j,i}$ 在被处理文章 A 中出现的次数.

由于在选择最佳切分时,统计值只是用于比较,且无论全局还是局部的统计值都只是似然估计值,故只具有相对意义,BUAASEISEG 在保持用于比较的词的统计值计算方式一致的前提下,灵活采用局部概率和全局概率对不同的切分进行比较(优先比较局部概率,当它们相等时则比较全局概率).另外,为了衡量切分出来的各个词的独立性,我们引入转移数(transition)的计算,即一个词 $w_{j,i}$ 转移到它的后续词 $w_{j,j}$ 的概率的总和,这里,$P(w_{j,i}w_{j,j})$ 是计算的局部概率,即在被处理文章 A 中出现的频率.

$$T(w_{j,i}) = \sum_{w_{j,j}} P(w_{j,i}w_{j,j}) \tag{6}$$

2 系统实现

2.1 分词流程

首先面临选词的问题,即选哪些词的序列进行切分评估.常用的 n 元切分总是要假设一个小的、并不实际的词长度的上界,为识别尽可能多的新词,我们提出采取二元迭代法进行 2^n 元词的选取.即首先以二元切分作为候选,筛选后,合并选取的二元词,把它们作为字进入下一轮迭代.如此反复,直到规定的迭代次数或没有符合要求的二元词出现(迭代收敛)为止.这种方法可以不设上界或设一个很大的上界.如:4 次迭代就可以发现 16 个字的词(一般汉字的词长在 8 个字以下),迭代收敛速度也很快(一般 3 次左右就收敛了).

系统分词的基本流程如图 1 所示.首先,对文本流用有限自动机(FSA)进行预处理,识别其中有明显特征的中英文数字(包括基数词、序数词、分数、小数)、域名、日期等.然后,BigramFilter 对 FSA 的输出进行过滤,并进行词频统计和候选词选择.BigramFilter 的输出又被输入到另一个 BigramFilter 中,如此反复不断地迭代,直到没有二元组可选为止.再次,基于前述分词结果进行一次最大匹配进行查漏补缺.最后,对分词后的文本进行倒排表索引.索引后的数据由于包含词和词在测试集中出现的频率,又可以作为新的词频词典来使用.若需要更高的准确率,可以用这个由测试集产生的新词频词典对测试集进行新一轮的分词流程,从而形成逐级递进

(bootstrapping)的增强学习.由图 1 可以看到,除了数据输入,BUAASEISEG 还需要一个词频词典或一个已分词的索引、中文姓氏列表、停词列表以及量词列表.我们没有限定名字的列表,是因为当代人总是倾向于起不重名的名字,任何训练集都不能包含全部姓名组合.

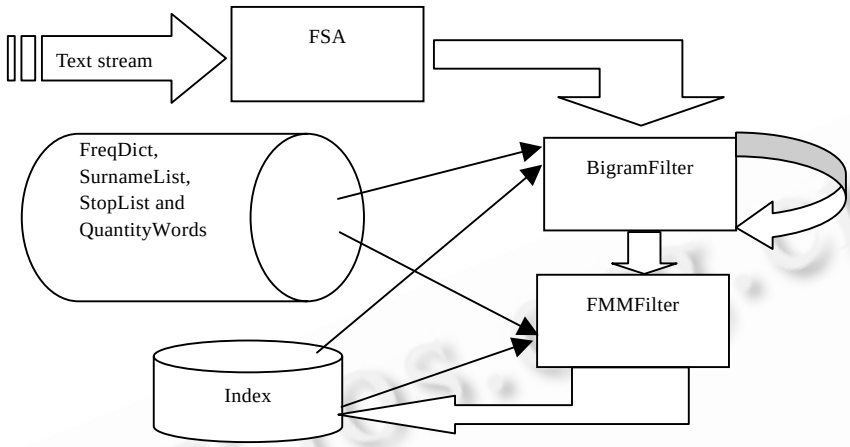

Fig.1 Word segmentation flow of BUAASEISEG

图 1 BUAASEISEG 分词流程

2.2 词频词典的准备

词频词典的准备不像标注训练集的准备那样需要耗费大量的人力及时间.我们用程序整合了 CDWS[5]统计的结果,包括拼音输入法中的词.从《人民日报》1998 年 1 月份标注语料中获得的词以及网上散布的一些其他小型词频表.词频表以词为索引,其出现的频率为值.词频表中最小词频大于后面将提到的发现新词的词频阈值.即收录到词频表中的词总是会作为候选词,这就要求词频表中没有错误的词.另外,我们删除了词频表中的中文数字、日期、序数词等,这样可以缩小词频表,因为这些有明确特征的词很容易被 FSA 找出.

2.3 分词算法

前文已经提到,我们利用反复调用 BigramFilter 来选择最长为 2^n(n 为调用次数)元的词,这种按照长度硬性拆分的方法必然会包含大量无意义的组合.我们的分词算法主要解决如何从中挑选出有意义的组合以及如何消歧的问题.BigramFilter 会对每次输入进行二元词频统计并计算转移(transition)次数.首先,选取所有和词频词典匹配的词.然后,根据前文的模型推导(式(5)),根据动态规划的思想,利用统计值从前到后一步步进行筛选.即

$$\sum_{i=1}^{m}\ln P(w_{j,i})=\left[\sum_{i=1}^{m-1}\ln P(w_{j,i})\right]+\ln P(w_{j,m}),\ 0<m<m_j.$$

对二元切分来说,句子 $S=C_1^n$ 在第 1 次迭代中的候选词 $CandiateBigram=\{C1C2,C2C3,C3C4,\ldots,Cn-1Cn\}$,通过比较 $P(C1C2)$和 $P(C2C3)$以及比较按式(6)计算的 $T(C1C2)$和 $T(C2C3)$,可以决定是选 $C1C2$ 还是 $C2C3$,依此类推.但这样依旧带有盲目性,没有包含任何启发信息,从而造成迭代中过多的比较.由于词频词典并不难构造,我们在词典的基础上进行统计分析,就能在很大程度上提高效率.在词典匹配的基础上,对于有歧义部分或非词典用词,我们定义了词频阈值和转移阈值.定义词频阈值是为了加快速度,尽快去掉无关的结果;转移阈值则决定了此 Bigram 的相对独立性.经过实验,我们把阈值都定义为 2 就能达到很好的效果.即此词至少出现两次,且有此词随后至少有两种转移.

由于使用基于二元(Bigram)的迭代,前一次的 Bigram 被当作后一次的 Gram.这样,筛选的梯度也是呈指数级增长的,解决了 n-gram 全切分所带来的组合爆炸问题.采用 Bigram 的迭代还避免了构造全切分网络,它每次只在两条路径中寻找更优的路径,直到走到终点,所走的路径就是局部最优的了,且迭代后可趋向于全局最优.

在筛选的一系列候选 Bigram 中,相互可能存在交叉歧义.我们通过比较前后两个词的词频和转移次数来决定是拆还是合.语义往往具有局部特征,同一篇文章中的词会集中表达相同或相似的语义.某些在训练集中少见

词的组合在特定语境下反复出现,反而更有代表性.因此,我们决定优先采用局部概率,其次采用全局概率.采用全局概率是不具备适应性的.例如:对于“曹勇刚才出来啊,曹勇刚太慢了!”这句话,意义很明确.但由于 ICTCLAS 训练语料中没有出现过曹勇刚或频率过少,它会把这句话分为:“曹 勇 刚才 出来 啊,曹 勇 刚 太慢 了”.而 BUAASEISEG 优先采用局部的统计,按照前述的规则,在第一次迭代时,由于“曹勇”的转移次数<“勇刚”的转移次数,故“曹勇刚”被拆分为“曹 勇刚”,第 2 次迭代时, “曹勇刚”的出现次数和转移次数都是 2,故“曹勇刚”被组合.当前后 Bigram 都存在时,会比较 3 个 Bigram 的局部概率和全局概率,以确定中间的 Bigram 是否拆分.如“郑重要求”会被拆分为“郑重 要求”,而不是“郑 重要 求”.而“研究生命苦”在“研究生命苦,研究生很累”中会被切分为“研究生 命苦”;在无上下文或“生命”和“研究”出现多的上下文中会被拆分为“研究 生命 苦”.这样就使得 BUAASEISEG 具有了很强的适应性.虽然对于仅在上下文中出现一次的新词或歧义,BUAASEISEG 也不能全部正确识别.但是,反映主题的新名词往往会在上下文中多次出现,因此,它的这个优势在实际使用时就十分明显了.

2.4 后处理

上一节介绍的只是纯基于统计和词频词典的算法.这样的结果对于识别新词已经非常有效了.但由于某些介词、代词、连词、助词等虚词常常和普通名词紧密结合,有较高新词识别率的基于统计的方法往往把它们切分到一起,我们的算法也不例外.构造停词表是对这一问题的有效弥补.但并不能发现停词就武断地切断 Bigram,因为某些新词可能包含停词,所以仍需依据此停词和其他部分的结合情况进行判断.设在某轮迭代中的二元模型由停词(SW)和另一部分(Ci)组成,P_L 代表局部概率.我们提出 3 条经验规则:(1) 优先切分——当且仅当 $P_L(SW\ Ci)=P_L(Ci)=P_L(SW)$时不予切分;(2) 中度优先——当且仅当 $2\times P_L(SW\ Ci)<P_L(Ci)+P_L(SW)$时进行切分;(3) 优先组合——当且仅当 $2\times P_L(SW\ Ci)\le P_L(Ci)$或 $2\times P_L(SW\ Ci)\le P_L(SW)$或($P_L(SW\ Ci)<P_L(SW)$且 $P_L(SW\ Ci)<P_L(Ci)$)时进行切分.

这 3 条规则是经验规则,各有偏向.如“比”是一个停词,但“比尔”就是一个词.对于规则 1,只有没有“比”字单独出现的文中,“比尔”才能被识别.对于规则 2,要求“比”和“尔”的次数总和超过了“比尔”的次数的两倍才进行拆分.对于规则 3,只有在包含“比尔”的文章中,“比”或“尔”的次数超过或等于“比尔”的次数的两倍,或者“比”和“尔”的次数都超过了“比尔”的次数,“比尔”才会被拆开.在面向信息检索的应用上,我们尽可能保证检索的准确度,也就是宁缺勿滥.所以我们采用了第 3 条规则处理停词,它适合停词表规模不大且包含的都是常用词的情况.规则 2 适合于设置了大量可能与新词冲突的停词时的情况.而采用规则 1 可以获得较高的召回率.本文后面的评测是基于规则 3 的结果.在人力资源允许的情况下,采取人工标注的方法经过训练来获取常用词的切分边界会更加理想,ICTCLAS 就采用了这种方法.

很多文章的人名仅出现一次,因此我们需要对人名进行特殊处理.通过对 GATE(http://gate.ac.uk)的姓氏表的增删,我们构造了姓氏表.姓氏是相对稳定的,但有些姓氏往往也是量词,如“年、元”等.所以在切分时,首先要识别量词,然后再识别姓氏.我们根据朱学锋等人总结的现代汉语量词[6]构造了量词表.为了提高效率,仅在怀疑是人名时才去判断是否是量词.当确信是姓氏后,会在 FMMFilter 中再根据它随后的词的长度和词频,以及它的转移数和转移的词(如摄、校)进行进一步判断.由于很多文章的作者都被明确分割,故此法对文章标题下以及参考文献中的作者姓名几乎全部能够识别(没有纳入姓氏表的例外).

3 结果评测及相关工作比较

由于各种判定标准以及评测数据集的不一致,分词的评测很难保证完全公平.选用测试集本身就带有不公平性.另一个问题是分词标准的不统一性,有些标注语料自身的标准就不一致[7],造成自动比较的困难,因此大批量的比较不太现实.如:我们为了检索的灵活性,倾向于把数词和量词分开,而 ICTCLAS 倾向于按照《人民日报》的标注方法把年月日和前面的数字放在一起.ICTCLAS 把姓和名分开,而我们需要把姓名放在一起,因为搜索“张勇刚”和“曹勇刚”是两个不同的查询.由于 ICTCLAS 是目前能够免费获取的最好的分词系统,区别于以最大匹配为基准进行比较的其他分词系统,我们以 ICTCLAS 为基准进行对比评测.这样,可同时看出

BUAASEISEGR 的长处和不足.

我们的目标是借助有效的分词系统让计算机更好地处理常用的信息.因此在语料的选择上,我们选择了主要的知识来源——新闻和科技文献.新闻类文章是按顺序取自 2005 年 7 月 15 日的 5 篇新浪网头条短篇新闻(编号为 1~5);科技文献是随机取的 NASAC2004 的 4 篇投稿论文(编号为 6~9).评测结果见表 1.评测过程是:由程序自动剔除非文字符号以及两者分词一致的部分(假定它们都对),然后对照原文人工分析不一致的部分.在比较差异时,取正确切分的词的数目作为计数.如: “试开采权”被 ICTCLAS 分为 3 个词,仍只认为它分错了一个词.对于前述不一致的分词规范部分以及纯粹由词典造成的差异,认为两者都对.

Table 1 Comparative evaluation between BUAASEISEG and ICTCLAS

表 **1** BUAASEISEG 和 ICTCLAS 的分词对比评测

No.	Num of characters	Num of words	Num of OOV words	Precision of BUAASEISEG (%)	Precision of ICTCLAS (%)
1	467	218	14	97.71	93.12
2	514	267	8	93.63	96.63
3	859	383	8	92.69	95.56
4	598	306	5	97.06	97.39
5	538	216	19	99.54	90.74
6	3 926	2 097	200	97.23	89.84
7	5 239	2 407	313	97.84	85.13
8	4 003	1 923	246	96.46	83.26
9	2 309	1 423	51	97.19	95.22
News	2 976	1 390	54	96.13	94.69
Papers	15 477	7 850	810	97.18	88.36
Both	18 453	9 240	864	96.66	91.53

从文章 6~9 可以明显看到,对于上下文丰富、新词很多的几千字的论文,BUAASEISEG 有着明显的优势;.从新闻 1、新闻 5 我们可以看到,由于上下文的帮助,BUAASEISEG 在识别新词的能力上,有着比较明显的优势;从对新闻 2、新闻 3 的分词结果看,ICTCLAS 在词法分析上的确具有不错的表现,它对介词短语和动词短语的分割相当出色,而 BUAASEISEG 由于新闻中大量重复出现类似于“通告称”,“向中国”这样的词语而造成较多切分失误(因为我们不把动词作为停词,“向”既是姓又是停词,且我们对停词采取优先组合的策略).由于人们倾向于检索非平凡的知识,而平凡的知识由于过多,缺少一部分对检索的整体效果影响不大,故这些常用短语对主题检索的影响甚微.也需要指出,由于采取了二元迭代局部优化策略,在对于某些只出现一次的三元词组(比如“发言人称”),BUAASEISEG 会出现切分错误,因为“发言”和“人称”都是词,而文中“称”只出现在“人”后面.因此,它会被拆分为“发言 人称”.

另外,BUAASEISEG 没有引入外文姓名的识别,ICTCLAS 对外文译名识别率较高,为它带来一定优势.但 ICTCLAS 对中文分数没有识别能力,它将“十一分之二”分为“十一分 之二”.在歧义消解方面,ICTCLAS 也存在明显的问题,如“预定义”和“存储量”分别被它分成了“预定 义”和“存 储量”.

综上所述,我们可以看出两个分词系统各有优势.BUAASEISEG 倾向于长词切分,更适合于进行主题发现,可用于提高搜索引擎的准确度,使用以它进行索引的搜索引擎,用户不会被引言中所述的一些无关的结果所困扰;而 ICTCLAS 则更适于词法分析,可进行细粒度切分,并能标注词性,特别是对于常用语的切分十分有效,但它对各领域的新词的识别能力以及词义消岐能力十分有限.因此,从总体上看,BUAASEISEG 倾向于基于语义的切分,比 ICTCLAS 部分基于语法、部分基于语义的切分更加合理.

微软亚洲研究院的 SCM[7]和 ICTCLAS 的思路类似.都考虑到了对式(1)进行转换,构造相应的隐马尔科夫链,从而达到同时解决分词和词性/词类标注的效果.因此,两个系统异曲同工,均可以采用成熟的 Viterbi 算法对问题进行求解.同样,SCM 和 ICTCLAS 虽然实现细节各异(体现在角色或类的定义以及对新词的发现规则/模板的定义上),但是按照这种方式选取的词序列具有相同的局限性,即受限于被标注的语料的覆盖面.训练语料的标注需要大量人力,很难做到全面覆盖.虽然 SCM 提出了逐级递进的方法来弥补训练语料的不足,但文中所提的逐级递进的方法需要人的持续参与,且主要是为了提高标注的准确率并进行歧义标识,而无助于发现训练集

中未出现的未知词(如:小波分析).本文的分词算法不需要繁复的训练语料标注过程,是一种非监督方法,可进行跨领域的未知词识别.

某些基于统计的分词系统对新词识别率也较高,如基于互信息、熵的词语提取算法[9]以及基于上下文的二元词切分[10]等.但它们在候选词的选取上要么依赖于词典,要么采用简单的二元切分,且对词频较小的词无能为力(因为阈值过大且没有对特殊词的处理),故也存在较大局限性.BUAASEISEG 在切分上提出了迭代的思想,从而拥有了切分长词的能力.并且,它利用了词频词典并引入了对特殊词的处理,因此即便在没有上下文的情况下也能完成基本切分.

为了增强分词系统的自适应性,高建峰等人提出了基于线形模型、根据不同拆分标准产生不同输出的方法,取得了良好的效果[11].这种方法从语料准备到系统实现,在空间和时间上都显得过于复杂,虽然实验结果良好,但其实用性仍受到一定限制.并且,这种分词系统对领域的适应性仍依赖于训练集的选取,由于领域的数量大且其划分也具有很多不确定性,很多文章可能是新兴领域或跨领域的,要对测试集自动区分领域并训练适当的分词器十分困难.相对来说,本文算法能在非监督的情况下进行优化选择,自身具有广泛的领域适应性,比较适合于领域混杂的网络文档,是一种简单而有效的方法.

最后,在效率和稳定性上,我们也进行了对比.为了有较强的可移植性,BUAASEISEG 全部使用解释型语言 JAVA 实现,它的切分速度约为使用 C++语言实现的 ICTCLAS 的一半(对于相同的程序,一般基于 JAVA 实现要比 C++的慢 10 倍左右[8],由此可知,从算法的角度来说,BUAASEISEG 算法的效率更高).在稳定性方面,ICTCLAS 开源版在处理包罗万象的网页文本过程中出现了严重的内存泄漏,在对约 600 万个网页的数据集 CWT100G(http://www.cwirf.org/SharedRes/Tool/CWT100g.html)进行分词时,处理完 30 多万个网页就能让拥有 1G 内存的机器内存溢出.而 BUAASEISEG 能够很顺利地处理完所有网页.

4 总结和展望

通过理论分析和实验,我们可以看到,BUAASEISEG 是一个具备较强的跨领域新词识别能力的分词系统.本文首次提出了基于二元迭代的切分方法(最长可以切分长度为 2^n 的词,n 为迭代次数),并提出了将搜索引擎索引数据作为词频词典实现逐级递进的分词的方法(第 1 次采用通用词频词典分词并索引,以后再利用索引作为输入再分词,再索引,如此循环往复,直至结果不再变化).BUAASEISEG 有着灵活的系统结构,在针对有上下文的文章的分词评测中,有着优良的表现,它具有很高的稳定性和可扩展性,是非常适合搜索引擎使用的分词系统.值得强调的是,不同于以往的中文分词系统,本系统不仅针对中文,对于特定含义的英文词组或短语也同样具有识别能力.目前,BUAASEISEG 对英文的识别主要是利用了前述分词算法中的统计模型,除了停词表以外,并没有为它进行其他语言相关的特殊处理.因此,它能发现词频较高且相对独立的英文词组,如“network tutoring”(网络教学).不仅如此,在 BUAASEISEG 的基础上进行中英文主题提取也是十分有效的(本文的关键词就主要来自于对本文的主题提取的结果),限于篇幅,有关内容将另文论述.以后的工作包括:引入对在上下文中出现频次少的外文译名的识别能力;对特殊中文语法现象的处理;对不合理切分的进一步排除以及对效率的进一步提升等.

致谢 感谢沈旭昆教授提供 CDWS 的词频词典,感谢李诺、吴安怡同学参与繁复的分词结果人工评测.

References:

[1] Foo S, Li H. Chinese word segmentation accuracy and its effects on information retrieval. Information Processing and Management, 2004,40(1):161−190.

[2] Zhang HP, Yu HK, Xiong DY, Liu Q. HHMM-Based Chinese lexical analyzer ICTCLAS. In: Proc. of the 2nd SigHan Workshop. 2003. 184−187.

[3] Su KY, Chaing TH, Chang JS. An overview of corpus-based statistics-oriented (CBSO) techniques for natual language processing. Computational Linguistics and Chinese Language Processing, 1996,1(1):101−157.

[4] Zhang HP, Liu Q. Model of Chinese words rough segmentation based on N-shortest-paths method. Journal of Chinese Information Processing, 2002,16(5):1−7 (in Chinese with English abstract).

[5] Liang NY. CDWS: A word segmentation system for written Chinese texts. Journal of Chinese Information Processing, 1987,1(2): 101−106 (in Chinese with English abstract).

[6] Zhu XF, Wang H. Classification of modern Chinese quantity suffix and noun. Technical Report, 1994 (in Chinese with English abstract). http://www.icl.pku.edu.cn/icl_tr/collected_papers/chinese/collection-2/yyy23.htm

[7] Gao JF, Li M, Huang CN. Improved source-channel models for Chinese word segmentation. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. 7−12.

[8] Giles JT, Wo L, Berry MW. GTP (general text parser) Software for text mining in statistical data mining and knowledge discovery. In: Bozdogan H, ed. Boca Raton: CRC Press, 2003. 455−471.

[9] Chang JS, Lin YC, Su KY. Automatic construction of a Chinese electronic dictionary. In: Yarowsky D, Church K, eds. Proc. of the 3rd Workshop on Very Large Corpora. 1995. 107−120.

[10] Dai YB, Khoo SGT, Loh TE. A new statistical formula for Chinese word segmentation incorporating contextual information. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 1999. 82−89.

[11] Gao JF, Wu AD, Li M, Huang CN, Li HQ, Xia XS, Qin HW. Adaptive Chinese word segmentation. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. 2004. 21−26.

附中文参考文献:

[4] 张华平,刘群.基于 N-最短路径方法的中文词语粗分模型.中文信息学报,2002,16(5):1−7.

[5] 梁南元.书面汉语自动分词系统——CDWS.中文信息学报,1987,1(2):101−106.

[6] 朱学锋,王惠.现代汉语量词与名词的子类划分.技术报告,1994. http://www.icl.pku.edu.cn/icl_tr/collected_papers/chinese/collection-2/yyy23.htm

曹勇刚(1977－),男,湖南长沙人,博士生,主要研究领域为知识/内容管理,文本挖掘,软件工程.

金茂忠(1941－),男,教授,博士生导师,主要研究领域为编译技术,软件工程.

曹羽中(1978－),男,硕士生,主要研究领域为文本挖掘,软件工程.

刘超(1958–),男,教授,博士生导师,CCF 高级会员,主要研究领域为软件工程.