一种基于合并策略的机构名称切分方法

钟 锋,罗燕京,杨 曦,李 虎
(北京航空航天大学 计算机学院,北京 100088)

摘 要:在就业招聘信息搜索系统中,如何正确切分机构名是一个非常重要的问题。在对机构名的组成结构进行了深入研究的基础上,提出了机构名的构成规则,建立了用于机构名切分的专有词典,并定制了一个基于合并策略的未登录词识别方法。本系统与海量分词系统进行了对比实验,实验表明,针对机构名切分这个特定领域,文中系统有更好的切分性能。在封闭测试中未登录词识别的准确率可以达到97.26%,召回率可达96.77%。

关键词:中文分词;机构名切分;1-最短路径算法;未登录词识别
中图分类号:TP391.12 **文献标识码**:A **文章编号**:1673-629X(2008)05-0012-03

An Organization Name Segmentation Approach Based on Combination Strategy

ZHONG Feng, LUO Yan-jing, YANG Xi, LI Hu
(School of Computer Science and Engineering, Beihang University, Beijing 100088, China)

Abstract: Organization name segmentation plays an important role in employment information retrieval system. Based on complete research of the organization name composition, the relevant structural features and domain dictionary were obtained. And also a combination approach is presented for unknown words identification in this paper. Experimental results show that the performance of the new system is better than several state-of-the-art systems in this special area. The experiment achieved 97.26% precision and 96.77% recall by close test.

Key words: Chinese word segmentation; organization name segmentation; one-shortest paths algorithm; unknown word identification

0 引 言

中文自动分词一直是中文信息处理技术中最为基础的课题,其主要应用于信息检索、汉字的智能输入、文本校对、自动摘要等很多方面[1]。目前,自动分词的基本算法主要分为两大类:基于词典的分词方法和基于统计的分词方法[2]。具体应用时的算法则是二者不同程度的组合[3]。

文中所要解决的机构名称切分问题(如"皇星珠宝公司"切分为"皇星//珠宝//公司")来源于"教育部毕业生招聘信息搜索系统"项目。对机构名称切分的目的是满足用户以公司名检索职位信息这一需求。

1 分词系统设计基本思想及目标

形式上,机构名称是由一个或一个以上的词加上表示机构称呼的名词(如"公司"、"集团"等)组成的。前者是后者的修饰语,即定语,后者是中心语[4]。如果把机构名称各部件分解归类,发现它们一般属于以下类型:

(1)地名。如:"北京空间装饰有限公司"(北京)。

(2)行业、专业及生产经营对象。如:"北京市迪克斯装饰集团"(装饰)。

(3)专有名、人名。如:"北京博纳电子有限公司"(博纳)。

(4)序数词。如:"中交第四航务工程勘察设计院"(第四)。

(5)机构称呼词(中心语)。如:"北京空间装饰有限公司"(有限公司)。

通过对以上机构名称各种部件词的归类分析,可以得到全称机构名称的一般性构造规则[4]:

{<地名>|<序数词>|{<人名>|<专有名>}|{<行业、专业>|<生产经营对象>}<机构称呼词>

在机构名称的组成各部件中,类型(1),(2)和(5)

收稿日期:2007-08-11
作者简介:钟 锋(1983-),男,山东滕州人,硕士研究生,研究方向为中文分词、中文信息检索;罗燕京,副教授,研究方向为中文信息处理、软件自动化测试。

属于普通词,可以通过训练语料库得到大部分并收入词典中[5]。而序数词有一定的规则和模式,可以通过正则表达式识别。类型(3)中的词是机构名称中的特征词,应作为新词识别,是机构名切分中需要解决的关键问题和难点。

2　具体设计与实现

文中采用"机械分词粗分+新词识别+歧义消除"这一流程进行处理(如图 1 所示)。

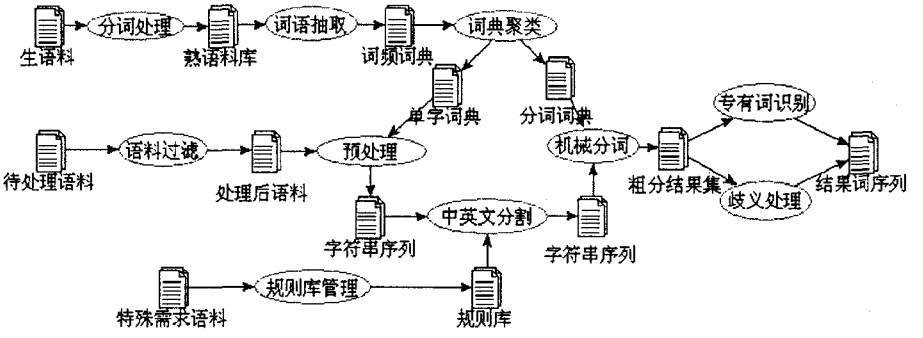

图 1　分词处理流程

2.1　词典生成

从含有 229188 机构名称库中,采用中科院的 ICTCLAS 实验版本进行切分和标注,辅助生成词典。然后人工整理和补充自动生成的词典(包含词频)。最终版本的词典只包含 7205 个词,高频词非常集中。另外需要说明的是词典中不含单字,如果切分结果集 > 1(说明有交叉性歧义)需要用到单字的词频,则默认为 1。这主要是因为根据对语料库分析,机构名中单字成词的概率非常小。

2.2　预处理与中、英文混合切分

这一步主要是根据标点、停词表、规则表达式,把字符串分解为不同类型片段,然后依据类型分别进行处理。在这里主要用到的是正则表达式分割[6]。主要有两个:

(1)对英文单词、数字的识别。

表达式:"[a-zA-Z0-9]+";例如"IBM 中国研究院"分割为"{IBM}+ENUM{中国研究院}+CJ",其中的"+ENUM"和"+CJ"都是片段的词性,"+CJ"表示为需要进一步切分的中文片段,其它的为词。

(2)对序数词的识别。

表达式:"第?[零一二三四五六七八九十|0-9]+";例如"东营市九六三计算技术研究所"分割为"{东营市}+CJ{九六三}+ORD{计算技术研究所}+CJ"。

2.3　1-最短路径算法及交叉性歧义处理

1-最短路径匹配(1-shortest path match)算法是先依据词典找出字符串中所有可能的词,然后构造词语切分的有向无环图,最后找出所有的最短路径(>=1)。具体的构造算法如下[7]:

设待切分字串 $S=c_1c_2\cdots c_n$,其中 $c_i(i=1,2,\cdots,n)$ 为单个的字,n 为串的长度,$n\geqslant 1$。建立一个节点数为 $n+1$ 的切分有向无环图 G,各节点编号依次为 $V_0,V_1,\cdots,V_n$。

通过以下两种方法建立 G 所有可能的词边:

(1) 相邻节点 V_{k-1},V_k 之间建立有向边 $<V_{k-1},V_k>$,边的长度值为 L_k,边对应的词默认为 $c_k(k=1,2,\cdots,n)$。

(2) 若 $w=c_ic_{i+1}\cdots c_j$ 是一个词,则节点 V_{i-1},V_j 之间建立有向边 $<V_{i-1},V_j>$,边的长度值为 L_w,边对应的词为 $w(0<i<j\leqslant n)$。

这样,待分字串 S 中包含的所有词与切分有向无环图 G 中的边一一对应。在计算时,假定所有的词都是对等的,为了计算方便,将词的对应边长的边长均设为 1。

例如,"深圳航建造价咨询公司"的切分有向图(如图 2 所示)有两条最短路径分别为:

a. 深圳 航 建 造价 咨询 公司(path 路径数组值:0,2,3,4,6,8,10);

b. 深圳 航建造 架 咨询 公司(path 路径数组值:0,2,3,5,6,8,10)。

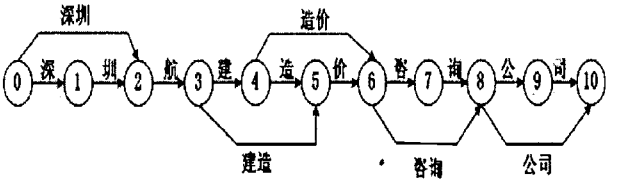

图 2　切分有向无环图实例

在这里采用基于一元独立性假设的统计模型进行歧异筛选。假设字符串 S,存在一种词的切分方式为 W。基于统计的切分法就是要求使得概率 $P(W\mid S)$ 最大的那个切分方式 W[4]。

文中采用的一元统计模型假定词与词之间是相互独立的,即有如下公式:

$$P(W\mid S)=\frac{P(S\mid W)\times P(W)}{P(S)}\approx P(W)$$

$$\cdots=P(w_1,w_2,w_i)$$

$$\cdots\approx P(w_1)\times P(w_2)\times\cdots\times P(w_i)\qquad(1)$$

其中在大规模语料库训练的基础上,根据大数定理,即:在大样本统计的前提下,样本的频率接近于其概率值。有:

$$P(w_i) = \frac{w_i \text{在语料库中出现的次数} n}{\text{语料库中总词数} N}$$

$$\cdots = k_i / \sum_{j=0}^{n} k_j \quad (2)$$

根据公式(1),(2)可以得到:

$$P(W) = \prod_{i=1}^{n} P(w_i) \approx \prod_{i=1}^{n} (k_i / \sum_{j=0}^{n} k_j) \quad (3)$$

在实际的计算中两边取对数有公式(4):

$$\ln P(W) = \sum_{i=1}^{n} [\ln P(w_i)] = \sum_{i=1}^{n} \ln(k_i / \sum_{j=0}^{n} k_j) = \sum_{i=1}^{n} [\ln k_i - \ln(\sum_{j=0}^{n} k_j)] \quad (4)$$

其中的常量是 $\ln(\sum_{j=0}^{n} k_j)$,故实际可以简化为计算 $\text{Max}(\sum_{i=1}^{n} \ln k_i)$。

2.4 专有特征词识别

未登录词是指分词词表中未收入的词,是汉语分词的难点问题。在机构名称的切分中,机构名称中的特征词就是未登录词,识别这些未登录词是整个分词系统的关键。

文中提出的识别策略是,首先通过 1 - 最短路径算法和正则匹配划分出那些确定的部分,然后再把没有归为这些确定类的词归为未登录词。这种方法简单易于实现,但是效果非常显著。

例如,"深圳航建造价咨询公司"(如图 2 所示),经过歧义处理的筛选, path 的值为"0,2,3,4,6,8,10"。切分结果和预想的一样,专有特征词"航建"被打散为单字,因此未登录词识别问题转化为对路径标识数组 path 的扫描、合并,即合并为"0,2,4,6,8,10"。在编码实现时,设计了 $O(n)$ 时间复杂度的算法,可以进行高效的合并。

2.5 详细设计与实现

在设计实现时,采用了 Lucene 分析器中使用的设计模式(如图 3 所示),即装饰(Decorator)模式,其主要优点是可以通过使用具体的装饰类以及这些装饰类的排列组合,创造各种不同的行为组合。在 Lucene 中,一个分析器主要包括分词器(Tokenzier)和过滤器(TokenFilter)两种部件,在设计中,主要使用了过滤器部件。各个分词器和过滤器构成了一个管道,文本在流过每个管道后,就是切分后的词流[8]。

具体到图 1 中的各处理步骤,预处理使用的是 Lucene 自有的 SimpleAnalyzer 分析器,它实现了按符号过滤分割待切分文本和英文的小写转换;图 3 中的 SplitFilter 实现中、英文混合切分,ZFCompanyFilter 机构名称中文部分的切分。为实现一些基于分词的其他功能(如输入的容错处理)可以实现一些其他的过滤器,然后连接上去就可以,保证了很好的灵活性和扩展性。

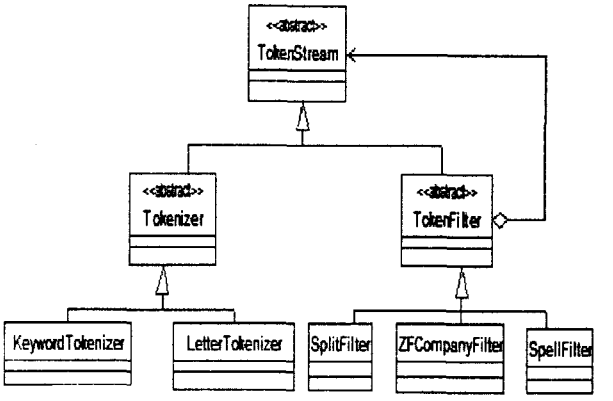

图 3 设计实现类图

3 测试与结果分析

文中所用的测试集是从中华英才网(http://www.chinahr.com)2007 年 3 月发布招聘信息的公司中抽取的 1000 条公司名称,在 P4 - 2.0GHz/1G 内存的机器上,分别对海量分词系统和文中系统进行测试。

针对机构名称的特点,文中考查的重点是未登录词的识别,即机构名称中专有特征词的识别。

目前,对未登录词识别普遍采用召回率(recall)和准确率(precision)两个评价指标,表 1 是两个系统评价指标的对比。文中系统在机构名分词方面取得了不错的效果,切词速度可以达到 78kb/s。

表 1 文中分词系统与海量科技分词系统对比

分词系统名称	未登录词总数	识别总数	正确识别数	召回率	准确率
文中系统	991	986	959	96.77%	97.26%
海量科技系统	991	791	732	73.86%	92.54%

注:召回率 = (正确识别的未登录词总数/文本中的所有未登录词总数) × 100%

准确率 = (正确识别的未登录词总数/识别出的未登录词总数) × 100%

对识别结果中的错误进行了分析,发现错误主要有以下几种类型:

* 词库中的词数量不足。

如:苏州工业园唯特利贸易有限公司→苏州//工业//园唯特利//贸易//有限公司(工业园)。此类错误可以通过进一步的增加训练集,扩充词量的方法解决。

* 组合性歧义问题。

如:深圳市天微电子有限公司→深圳市//天//微电子//有限公司。这种组合型的歧义是算法本身的特点造成的(1 - 最短路径算法),不容易解决。不过这种组合型歧义在所有的歧义中所占的比率很小,文中的

(下转第 18 页)

平均绝对误差(MAE):$MAE = \frac{\sum | e_i |}{N}$

均方差误差(MSE):$MSE = \frac{\sum (e_i)^2}{N}$

均方根误差(RMSE):$RMSE = \sqrt{MSE}$

平均相对误差(MAPE):$MAPE = \frac{1}{N} \sum | \frac{e_i}{s_i} |$

算法运行时间 RUNT:为预测算法开始到结束的时间间隔。

其中 e_i 为汇率实际值与预测值的绝对误差,s_i 为汇率实际值。$i = 1:16, N = 16$。

表 1 各项性能指标比较

各项性能指标	BP 神经网络	SVM 回归
平均绝对误差(MAE)	6.8750e-005	5.0000e-005
均方差误差(MSE)	8.1250e-009	6.2500e-009
均方根误差(RMSE)	9.0139e-005	7.9057e-005
平均相对误差(MAPE)	4.5280e-007	1.0062e-007
算法运行时间 RUNT	400s	≈1s

仿真结果表明,SVM 回归算法的预测效果良好,能很好地跟踪即时汇率的变化趋势,预测精度高。与 BP 神经网络相比 SVM 回归算法可以提供更好的泛化能力,且运行速度要快的多。

3 结束语

汇率预测的研究一直是国内外学者研究的热点。超短期汇率预测的研究对汇率日常交易具有十分重要的意义。针对目前流行的神经网络预测模型不能适应超短期预测的要求,文中将支持向量机回归应用于即时汇率的预测中,获得了良好的拟合与预测精度。但是如何更加合理地选择各模型中的参数仍需进一步研究。

参考文献:

[1] Maguire L P, Roche B, Mcginnity T M. Predicting a chaotic time series using a fuzzy neural network[J]. Information Sciences,1998,112:125-136.

[2] Vapnik V N. The nature of statistical learning theory[M]. New York:Springer-Verlag,1995.

[3] Grassberger P, Procaccia I. Characterization of Strange Attractors[J]. Phys Rev Let,1983,50(5):346-349.

[4] Gautama T, Mandic D P, Vanhulle M M. A differential entropy based method for determining the optimal embedding parameters of a signal[C]// Proc of the Int Conf on Acoustics, Speech and Signal Processing. Hong Kong: [s. n.], 2003:29-32.

[5] Schreiber T, Schmitz A. Surrogate time series[J]. Physica D, 2000,142(3-4):346-382.

[6] 崔万照, 朱长纯. 混沌时间序列的支持向量机预测[J]. 物理学报,2004(10):3303-3309.

[7] 黄巧玲,谢维波. 超短期汇率的预测研究[J]. 计算机应用, 2007(4):1009-1012.

(上接第 14 页)

统计结果为 2.1%,可以暂时不予处理。

* 地名后缀问题。

如:珠海市尼科希电子有限公司→珠海//市尼科希//电子//有限公司。解决问题的办法是建立地名后缀库,如"市"、"区"、"省"、"镇"等加入库中。

4 结 语

由于文中涉及切分需求的特殊性,目前的分词系统针对该领域分词效果不理想,所以文中定制了一个基于合并策略的未登录词识别方法,实验表明,在针对机构名切分这个特定领域,文中分词系统有更好的切分性能。另外,在一些短语类汉语切分中,该方法同样有效,前提是增加相关的关键词库[9]。

参考文献:

[1] 吴 栋,滕育平. 中文信息检索引擎中的分词与检索技术[J]. 计算机应用,2004,24(7):128-131.

[2] 朱巧明,李培峰,吴 娴,等. 中文信息处理技术教程[M]. 北京:清华大学出版社,2005:193-196.

[3] Lai B Y, Sun M S. Chinese word segmentation and part-of-speech tagging in one step[C]// Proceedings of International Conference: Research on Computational Linguistics. Taipei: [s. n.], 1997:229-236.

[4] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆,2000:4-10.

[5] Chen Keh-jiann, Chen Chao-jan. Knowledge extraction for identification of Chinese organization names[C]// In proceeding of the 19th International Conference on Computational Linguistics. Taipei:[s. n.],2002.

[6] Knuth D E, Morris J H, Pratt V R. Fast pattern matching in strings [J]. SIAM Journal on Computing, 1977,6(2):323-350.

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

[8] 林碧英,赵 锐,陈良臣. 基于 Lucene 的全文检索引擎研究与应用[J]. 计算机技术与发展,2007,17(5):184-186.

[9] 钟良伍. 基于中文机构名简称的检索方法研究 [D]. 北京:清华大学信息科学技术学院,2005.