一种引入动态词库更新的中文分词架构①

刘芳芳, 王 晶, 沈奇威

(北京邮电大学 网络与交换技术国家重点实验室, 北京 100876)
(东信北邮信息技术有限公司, 北京 100191)

摘 要: 针对互联网环境下新词出现和更新频率高的特点, 将机械分词与基于规则分词相结合, 提出一种动态更新词库的中文分词架构. 本架构给出了新的词典设计结构及歧义处理规则, 并将统计学中的互信息概念运用到新词判定环节. 实验表明本文提出的中文分词架构具有较高的准确率和良好的适应性.

关键词: 中文分词; 最大词长; 歧义处理; 互信息; 未登录词

Framework for Chinese Word Segmentation with Dynamic Updating Dictionary

LIU Fang-Fang, WANG Jing, SHEN Qi-Wei

(State Key Lab of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)
(EBUPT Information Technology Co. Ltd., Beijing 100191, China)

Abstract: Aiming at new word appearing and updating frequently in the Internet, through combining mechanical segmentation and rule-based segmentation, a framework for Chinese word segmentation with dynamic updating dictionary is provided. We improve a new dictionary design structure and ambiguity processing rules. Also the concept of mutual information in Statistics is used in determining the new word. The experiment shows that the framework for Chinese word segmentation proposed in this paper has high precision and is adaptive in different domains.

Key words: Chinese word segmentation; maximum word length; ambiguity processing; mutual information; unlisted words

1 引言

在中文信息处理的众多领域中, 包括信息提取、信息检索、机器翻译、语音识别以及针对 Web 数据挖掘应用提出的诸如文本分类、聚类、发现关联规则等技术中, 中文分词是基础研究课题, 也是中文信息处理的瓶颈问题[1,2]. 众所周知, 英文以空格作为自然的分隔符, 而中文由于继承自古代汉语的传统, 词语之间没有分隔. 中文只是字、句和段能通过明显的分界符来划界, 唯独词没有形式化的分界符.

汉语自动分词已被研究了 20 多年, 在这期间所提出的自动分词方法有 20 多种. 这些分词方法概括起来可分为两大类: 一类是基于词典和统计相结合的机械分词方法; 一类是基于规则的专家系统的分词方法. 不管是哪类分词方法, 都会有错误的切分. 如果在词语切分当中所运用的切分算法较好, 错误切分的句子数量将会大大减少, 再采用相应的消除歧义策略进行消歧, 就会进一步提高分词的精度.

2 互联网环境下的中文分词

目前切分精度较高且是开源版本的中文分词有中科院计算所的 ICTCLAS 分词系统和开源分词组件 IKAnalyzer, 前者 ICTCLAS 在“973”专家组评测中, 分词正确率高达 97.58%; 后者最新版本 IK2012 经小范围测试大概准确率在 90% ~ 95%之间. 但是将上述中文分词系统应用到互联网环境下, 对网络上的文章进行分词应用, 发现并不能达到预期的准确率. 测试结果见表 1. 测试数据随机摘自国内最大的中文门户网站搜狐网[3]和国内最大的生活消费指南网站大众点评网[4].

① 基金项目:国家自然科学基金(61072057,61101119,61121001,60902051);长江学者和创新团队发展计划(IRT1049);国家科技重大专项(2011ZX03002-001-01)
收稿时间:2012-10-09;收到修改稿时间:2012-10-31

表 1 互联网样本分词测试结果

	1	2
原文本数据	神舟九号飞船返回舱顺利降落在内蒙古中部主着陆场预定区域	地址：官渡区关兴路 75 号(宝海公园大门对面)
ICTCLAS 切分结果	神/舟/九/号/飞船/返回/舱/顺利/降落/在/内蒙古/中部/主/着/陆/场/预定/区域	地址/官渡区/关/兴/路/75/号/宝/海/公园/大门/对面
IK2012 切分结果	神舟/九号/飞船/返回舱/顺利/降落在/ 内蒙古/中部/主着陆/场/预定/区域	地址/官渡区/关兴/路/75 号/宝/海/公园/大门/对面
人的理解	神舟九号/飞船/返回舱/顺利/降落/在/内蒙古/中部/主着陆场/预定/区域	地址/官渡区/关兴路/75/号/宝海公园/大门/对面

对两种分词系统的准确率的统计结果见表 2.

表 2 互联网样本分词测试统计结果

	人的理解(词)	ICTCLAS		IKAnalyzer	
		准确词(词)	准确率	准确词(词)	准确率
统计	19	14	73.7%	11	57.9%

从上面的测试结果可以看出，互联网环境下，中文语言的使用习惯和造词能力发生变化，而机械分词依赖的词典不可能囊括互联网上出现的所有新词，所以造成对传统语言切分精度很高的中文分词系统，应用在互联网环境下时，切分精度急剧下降.

网络语言与传统语言相比有一些显著的新特性：新词出现和更新速度加快；网络语言的使用习惯口语化、简写化、不规则化，例如存在对新生事物或者机构团体名的简称，对店铺及菜肴的命名多追求新颖，这时如果按照传统汉语的规则和语法进行静态词典的匹配和切分就会产生误差，因此应用于开放互联网环境下的中文分词系统有待进一步改善.

3 一种引入动态词库更新的中文分词架构

针对互联网环境下特定领域的语言特点，在对分词算法的理解和对已有最大匹配分词算法的分析和研究的基础上，本文提出一种引入动态词库更新的中文分词算法，通过本算法，在进行传统中文分词的过程中，动态识别未登录词，同时引入判定规则，对符合条件的未登录词动态更新到中文分词所依赖的词典，提高中文分词算法的适应性和正确率. 经过实验证明，本文所提出的算法在互联网环境下对新词的识别和切分有较高的准确率.

3.1 中文分词过程

目前常用的自动分词系统大多采用机械分词为主，辅以少量的语法和语义信息的分析. 机械分词是基于字符串匹配的原理进行的，其优点是原理较为简单，易于在计算机上实现，时间复杂度也相对降低[5]. 本文同时采用了机械分词中常用的正向最大匹配法和逆向最大匹配法进行双向匹配，目的是有效识别分词中存在的歧义字段，并且改进了传统机械分词中的词典结构，本文提出的基于词库和机械分词与统计规则分词相结合的中文分词算法，有效提高了分词的正确率和对未登录词的识别效率.

本文的中文分词算法由输入、预处理、双向最大匹配切分、歧义处理、未登录词识别、输出等几部分组成，具体流程图见图 1 的虚框 1.

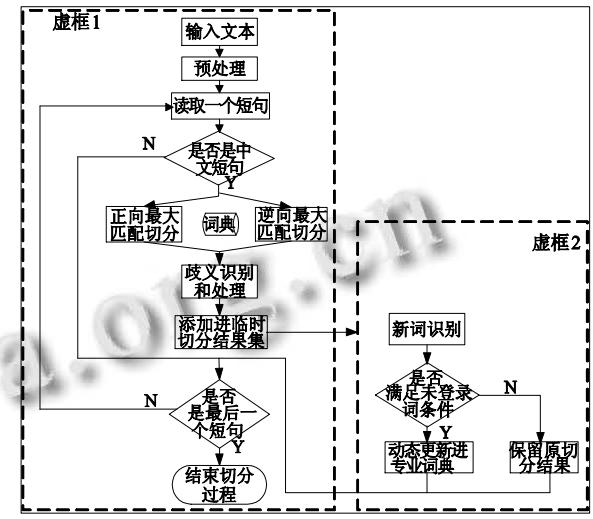

图 1 系统整体流程图

1) 词典

最大匹配分词算法有一个严重的问题是：最大词长的长度比较难于确定，如果定得太长，则匹配时花费的时间就多，算法的时间复杂度明显提高，如果定得太短，则不能正确切分更长的词，导致切分正确率降低[6]. 本文设计了一种可以在句子范围内匹配时动态更新最大词长的词典结构. 词典由一个首字词长索引表，一个尾字词长索引表和词典正文构成. 其

中首字词长索引表包含两项内容, 一个是词典正文中所有单词的首字, 另一项为词典正文中以该首字开头的最大词长. 尾字词长索引表也包含两项内容, 一个是词典正文中所有单词的尾字, 另一项为词典正文中以该尾字结束的最大词长. 两个索引表中的单字按照汉字在计算机中的机内编码排序, 查找某个单字时根据首字的机内码直接得到对应的最大词长, 时间复杂度为 O(1). 在进行最大匹配分词时, 则根据当前字符串序列的首字或尾字决定要截取的子字符串长度, 从而有效减少查找及匹配次数. 词典正文由 9 个按词长从 2 到 10 的基本词典文件构成, 每个子词典的单词字数相同且已经按照各个单词的机内编码排序, 查找某个单词时根据其机内编码二分查找词典文件, 提高查找速度.

词典分为基本词典和专业词典, 基本词典用于机械分词时的匹配词典, 专业词典用于动态更新词库时保存未登录词, 方便加载与卸载专业词典.

2) 预处理

预处理模块对要分词的一段文本, 利用标点、英文字母、数字、其它符号等分割成一个个长度较短的子串, 新词判定则只对包含中文字符的子串进行.

3) 处理中文短句

对中文短句采取双向匹配切分, 目的是通过对比两次切分结果, 寻找并处理歧义字段. 双向匹配切分即正向最大匹配切分和逆向最大匹配切分.

正向最大匹配切分的过程描述如下: 假设被处理材料当前字符串 S 序列的首字为 $SChar$, 查询首字词长索引表得到首字 $Schar$ 的最大词长为 m, 比较 m 与 S 的句长, 取较小值, 设为 n. 取被处理材料当前字符串 S 序列中的前 n 个字作为匹配字段, 查找分词词典, 若词典中有这样一个 n 字词, 则匹配成功, 匹配字段作为一个词被切分出来, 如果词典中找不到这样的一个 n 字词, 则匹配失败, 匹配字段去掉最后一个汉字, 剩下的字符作为新的匹配字段, 再进行匹配, 如此进行下去, 直到匹配成功为止. 有专家统计过, 理论正向最大匹配法的错误切分率为 1/169.

逆向最大匹配切分的分词过程与正向最大匹配切分相同, 不过是从句子末尾开始处理, 首先查询尾字词长索引表得到尾字的最大词长 m, 每次匹配不成功时去掉的是前面的一个汉字. 据专家统计其分词精度比正向最大匹配法要高, 其错误切分率是 1/245.

4) 歧义识别和处理

只含中文字符的短句经过双向匹配切分后, 对比正向最大匹配切分和逆向最大匹配切分的切分结果, 进入歧义识别和处理阶段: 如果两次切分结果中单词的数目不相同, 则按照长词优先选择切分数目少的作为切分结果; 如果两次切分数目相同则依次扫描比较两个切分结果中的每个单词, 同时记录切分结果差异字段. 如果扫描完毕, 差异字段为空, 则认为切分结果正确, 选择正向匹配结果作为切分结果, 否则, 对差异字段进行歧义处理.

通过对交叉歧义字段的考察, 发现其中 80%以上可以通过运用一条无需任何语言知识的“归右原则”(交叉歧义字段优先与其右边的字段成词)就可以获得正确切分, 这是因为在多数情况下汉语的修饰语在前、中心词在后, 因而“归右”好于“归左”, “归右原则”可以使机械分词的精度上升到 99.4%[7]. 同时参考文献[7], 对连续性交叉歧义和介词框型标志的交叉歧义进行处理. 本文通过对文献[7-11]中的典型交叉歧义字段的例句分析发现, “回归”比较“归右”处理后的差异字段的最后一个单词, 如果与差异字段正向匹配切分结果的最后一个单词相同则取正向切分结果, 如果与歧义字段逆向匹配切分结果的最后一个单词相同则取逆向切分结果, 如果都不同则按照长词优先取正向切分结果, 正确率可以达到 85%以上. 利用文献[9]中的统计数字, 词法歧义字段、句法歧义字段、语义歧义字段和语用歧义字段在语料中的出现次数和语料中所含汉字总数之比分别为: 0.766%、0.098%、0.031%和 0.016%. 本文提出的“归右再回归”的机械分词的精度为:

1-[(1-85%)*0.766%+0.098%+0.031%+0.016%]=99.74%.

5) 处理非中文短句

在此模块中正确识别并处理连续的字母、连续的数字及字母和数字的组合情况, 如字母间有连接符, 浮点数, 百分号, 电话号码, IP 地址等.

经过上述几个阶段的中文分词处理, 预处理后中文字符构成的短句则被切分为一个个单词的序列, 此序列则进入新词识别及词库动态更新过程.

3.2 动态更新词库过程

未登录词即未包括在分词词表中但必须切分出来的词, 它大致包含两大类: l) 新涌现的通用词或专业术语等; 2) 专有名词, 如中国人名、外国译名、地名、机构名(泛指机关、团体和其它企事业单位)等. 前

一种未登录词理论上是可预期的，能够人工预先添加到词表中(但这也只是理想状态，在真实环境下并不易做到)；后一种未登录词，由于可能的中文姓名、译名数量庞大和小地名太多、太杂不宜采用穷尽收录法．因此，只能寻找其它方法来处理，统计和规则相结合的方法是一种有效的识别方法．本文主要处理第二种未登录词．通过对大量语料的研究表明，90%的未登录词是以单字形式出现，因此，本文对于未登录词的识别主要是针对切分碎片，采用非结束标志单字向右“砌”和结束标志单字向左“拼”或右“拼”的规则寻找新词，规则主要基于两方面知识：未登录词的内部结构信息和上下文信息．前者是指未登录词本身的结构特征(如用字特点)，用来指示未登录词的存在并确定其有效性；后者是指未登录词所处的上下文环境，用于切分未登录词的边界并确定其类型．本文根据大量语料得到姓氏候选集、地名用字后缀集(如：路，村，镇等近 40 个)、机构名用词集合(如：广场，商城，大厦，宾馆等近 50 个)，这些标志字符集一方面用于新词本身边界的确定，另一方面用于新词是否可以作为未登录词及是否可以动态更新进词库的判定．

从中文分词过程后形成的单词序列开始，单词序列没有单字或者所有单字判断完毕结束．如果单词序列存在单字，则从第一个非停用词单字开始，如果不属于上述的标志字符集，则继续与后面的单字右“砌”，直到遇到非单字进入判定阶段；如果属于上述的地名标志字符集，则左“拼”一个非地名标志且非机构名结尾的单词，进入判定阶段，单字本身为地名标志字符且前一个单词为地名标志或者机构名则属于特殊情况，需要单做处理．人名识别改成右“拼”，过程类似．

上述算法会拼接出“的人”、“及在”等高频共现但非词的字串，本文特引用信息论中用来判断两个事件集合之间相关性的“互信息”概念来判定新词是否可以作为未登录词及是否可以动态更新进词库．结合新词在原文中的上下文信息，本文对互信息的概念加以修正，引入新的统计值：左互信息，右互信息和本词自有信息．左互信息定义为新词左侧是否出现对新词有识别意义的词汇，如对地名中新识别的街道名左侧是否出现“路”的标志字符．右互信息定义为新词右侧是否出现对新词有识别意义的词汇，如机构名“商城、公司等”．本词自有信息定义为识别出的新词对自己成为未登录词能否提供决定信息，如对地名而言是否是标志字符结束．在判定阶段，对每一个新词进行左/右互信息和本词自有信息的计算，满足条件的即被系统识别为未登录词，按词长动态更新进专业词典文件，更新保存词典文件，具体流程见图 1 中的虚框 2．通过对大语料集的新词学习过程形成某一领域的专业词典，在此基础上进行本领域的中文分词与信息处理，经过实验验证，本文提出的架构具有很高的准确率．同时本文采取专业词典可动态卸载的方式，不影响原基本词典对其他领域的分词效果．

4 测试结果及结果分析

实验文章来自大众点评网，所有分词均为自动分词结果．

原文是：西山区卢家营新村西华小区夏蓉里停车场对面．

切分结果：西山区/卢家营/新村/西华小区/夏蓉里/停车场/对面．同时识别的未登录词西山区，卢家营，西华小区，夏蓉里被正确更新并保存进专业词典．

在歧义消除方面我们使用一个比较经典的例子进行测试．原句是：“长春市长春节讲话．”正向最大匹配切分结果为：“长春市/长春/节/讲话”，逆向最大匹配切分结果为：“长春/市长 春节/讲话”，识别出歧义字段“长春市长春节”，“归右”处理后歧义字段最后一个单词为“春节”，与逆向最大匹配对歧义字段的切分结果相同，因此最终的分词结果取逆向最大匹配切分的分词结果．

根据本文对近 20 万条的昆明商家名称及地址的规模测试及统计，对未登录词识别的准确率达 95%，高于互联网环境下 ICTCLAS 系统的分词准确率．

5 结语

本文将机械分词及基于规则分词相结合，为正向最大匹配分词设计了新的可动态确定最大切分词长的词典结构，引入了新的歧义处理规则，基于切分过程产生的单字碎片，引入右“砌”左“拼”或右“拼”的规则寻找新词，同时采用了统计学中的方法进行判定，实现了词库可动态增长的中文分词架构．但如何利用上下文信息更好的处理组合型歧义和真歧义，如何将本文提出的架构更好的应用于更多的领域，还有待于更深层次的研究．

(下转第 50 页)

目前 E 系统在主流程(一阶流程)部分, 统一采用工作流进行调度和管理, 二阶流程则根据各子系统的需要, 采用轻度或者中度的流程与业务整合.

4 结语

专利审批过程的流程化特征, 使得采用流程管理模型进行调度和控制成为了一种需要. 工作流技术为这方面的需要提供了技术支撑和理论基础. 业务上复杂的流程性操作, 在工作流模型的应用下, 实现了清晰的分层, 明确的定义, 统一的过程记录, 集中的调度策略.

中国专利电子审批系统中的工作流引擎构建在 J2EE 平台上, 流程定义使用 XML 文件, 后台数据库使用 Oracle10g. 投入实际运行的工作流引擎稳定、灵活, 并为维护管理人员在专利数据分析、定位问题等方面提供了大量的帮助. 专利电子审批系统中工作流技术应用的成功经验, 可作为其它具备流程化特征的电子政务系统、企业管理系统的一个参考.

致谢 本文得到专利电子审批系统项目组的支持, 并得到张宇主任、陆新年处长和唐俊松处长的帮助和指导, 审稿人对本文提出了宝贵的修改意见, 谨致谢意.

参考文献

1 Workflow Management Coalition.The Workflow Reference Model[WfMC1003].WFMC TC00-103,1995.

2 罗海滨,范玉顺,吴澄.工作流技术综述.软件学报,2000,11(7):899−907 .

3 van der Aslst WMP. Three good reasons for using a Petri-net -based workflow management system. Cambridge, MA: Kluwer Academic Publishers, 1996: 179–201.

4 Loadmaster S. Petri net. 2012.http://en.wikipedia. org/wiki/Petri net.

5 屈婉玲,耿素云,张立昂.离散数学.北京:高等教育出版社, 2008.

6 赵文,胡文蕙,张世琨,王立福.工作流元模型的研究与应用.软件学报,2003,14(6).

7 OpenSymphony project. OSWorkflow, 2006, http://www. open-symphony.com/osworkflow.

8 Kuru B. Workflow. 2012. http://en.wikipedia.org/wiki/Workflow.

(上接第 103 页)

参考文献

1 张春霞,郝天永.汉语自动分词的研究现状与困难.系统仿真学报,2005,17(1):138−147.

2 Ni P, Liao JX, Wang C, Ren KY. Web information recommendation based on user behaviors. 2009 WRI World Congress on Computer Science and Information Engineering. 2009: 426–430.

3 http://www.sohu.com/.

4 www.dianping.com.

5 文庭孝,邱均平,侯经川.汉语自动分词研究展望.现代图书情报技术,2004,(7):6–10.

6 梁晓弘,杨文安.分词技术在信息处理中的研究综述.电脑知识与技术,2007(22):1100−1102,1117.

7 骆正清,胡上序,陈增武.一种改进的 MM 分词方法的算法设计.中文信息学报,1996,10(3):30−36.

8 梁南元.汉语计算机自动分词知识.中文信息学报,1990,4(2):29−33.

9 何克杭,等.书面汉语自动分词专家系统设计原理.中文信息学报,1991,5(2):1−14.

10 姚天顺,等.基于规则的汉语自动分词系统.中文信息学报,1990,4(1):37−43.

11 谭琼,史忠植.分词中的歧义处理.计算机工程与应用,2001(11):125−127,236.