国内中文自动分词技术研究综述*

奉国和[1] 郑 伟[2]
[1]华南师范大学经济管理学院 广州 510006 [2]河北北方学院理学院 张家口 075000

〔摘要〕认为分词是文本自动分类、信息检索、信息过滤、文献自动标引、摘要自动生成等中文信息处理的基础与关键技术之一,中文本身复杂性及语言规则的不确定性,使中文分词技术成为分词技术中的难点。全面归纳中文分词算法、歧义消除、未登录词识别、自动分词系统等研究,总结出当前中文分词面临的难点与研究热点。
〔关键词〕中文分词 分词算法 歧义消除 未登录词 分词系统
〔分类号〕G354

Review of Chinese Automatic Word Segmentation

Feng Guohe[1] Zhen Wei[2]
[1]School of Economics & Management, South China Normal University, Guangzhou 510006
[2]College of Science, Hebei North University, Zhangjiakou 075000

〔**Abstract**〕Word segmentation is one of the key technology for natural language processing such as text auto - classification, information retrieval, information filtration, document auto - index, summarization auto - generation etc.. Chinese word segmentation is difficult problem in word segmentation because of it's complexity and uncertain language rules in nature. This paper sums up the research comprehensively of Chinese word segmentation algorithm, disambiguation method, unknown word recognition, auto - segmentaion systems etc. and summarizes Chinese word segmentation's research difficult points and hot points today.
〔**Keywords**〕chinese word segmentation word segmentation algorithm disambiguation method unknown word recognition word segmentation system

中文分词是文本分类、信息检索、信息过滤、文献自动标引、摘要自动生成等中文信息处理中的关键技术及难点。经过广大学者共同努力,过去20多年中文分词取得可喜进步,黄昌宁、赵海[1]在四方面总结了取得的成绩。笔者利用CNKI全文期刊数据库,以"中文and分词"、"汉语and分词"、"自动and分词"等为检索条件,检索时段为1987年1月1日-2010年9月11日,进行篇名检索,经筛选分别得到相关研究论文214、191、165篇,通过文献归纳总结出该领域研究现状、研究内容、研究热点与难点,并展望其发展。

1 中文分词基础理论研究

中文分词理论研究可归结为:三种主要分词算法及组合算法研究、中文分词歧义消除、未登录词识别与分词与词性标注评测研究。

1.1 分词算法研究

衡量分词算法优劣标准是分词速度与精度,各种算法围绕精度与速度展开。目前分词算法很多,大致可归纳为:词典分词方法、理解分词方法、统计分词方法、组合分词算法。

1.1.1 词典分词方法

• 算法。词典分词方法按照一定策略将待分析汉字串与词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功,该方法需要确定三个要素:词典、扫描方向、匹配原则。比较成熟的几种词典分词方法有:正向最大匹配法、逆向最大匹配法、双向最大匹配法、最少切分等。实际分词系统,都是把词典分词作为一种初分手段,再通过各种其他的语言信息进一步提高切分的准确率。

词典分词方法包含两个核心内容:分词算法与词典结构,算法设计可从以下几方面展开:①字典结构改进;②改进扫描方式;③将词典中的词按由长到短递减顺序逐字搜索整个待处理材料,一直到分出全部词为止。

*本文系国家社会科学基金项目"自动文本分类技术研究"(项目编号:08CTQ003)研究成果之一。
收稿日期:2010-08-12 修回日期:2010-09-13 本文起止页码:41-45 本文责任编辑:杜杏叶

• 词典结构。词典结构是词典分词算法关键技术,直接影响分词算法的性能。三个因素影响词典性能[2]:①词查询速度;②词典空间利用率;③词典维护性能。Hash 表是设计词典结构常用方式,先对 GB2312 - 1980 中的汉字排序(即建立 Hash 表),然后将其后继词(包括词的属性等信息)放在相应的词库表中。

孙茂松等[3]设计并实验考察了三种典型的分词词典机制:整词二分、TRIE 索引树及逐字二分,着重比较它们的时间、空间效率。姚兴山[4]提出首字 Hash 表、词次字 Hash 表、词次字结构、词 3 字 Hash 表、词 3 字结构、词 4 字 Hash 表、词 4 字结构、词索引表和词典正文的词典结构,该结构提高查询速度,但增大存储开销。陈桂林[5]等介绍了一种高效的中文电子词表数据结构,它支持首字 Hash 和标准的二分查找,且不限词条长度,并给出利用近邻匹配方法来查找多字词,提高了分词效率。目前文献看,围绕词典结构提高分词性能的主流思想是设计 Hash 表,表数目随结构不同而不同,数目越多,空间开销越大,但查询速度也相应提高,具体设计需要在时间与空间之间权衡。

1.1.2 *理解分词方法* 基本思想是分词同时进行句法、语义分析,利用句法信息和语义信息来处理歧义现象,理解分词方法需要使用大量语言知识和信息。

• 人工智能技术。人工智能技术主要包括专家系统、神经网络和生成 - 测试法三种。分词专家系统能充分利用词法知识、句法知识、语义知识和语用知识进行逻辑推理,实现对歧义字段的有效切分。何克抗等[6]深入分析了歧义切分字段产生的根源和性质,把歧义字段从性质上划分为四类,并给出消除每一类歧义切分字段的有效方法。王彩荣[7]设计了一个分词专家系统的框架:将自动分词过程看作是基于知识的逻辑推理过程,用知识推理与语法分析替代传统的"词典匹配分词 + 歧义校正"的过程。神经网络摸拟人脑神经元工作机理设计,将分词知识所分散隐式的方法存入神经网内部,通过自学习和训练修改内部权值,以达到正确的分词结果。林亚平[8]、尹锋利等[9]用 BP 神经网络设计了一个分词系统,进行大量仿真实验,取得不错分词效果。

采用神经网络与专家系统的人工智能分词算法与其他方法相比具有如下特点:①知识的处理机制为动态演化过程;②字词或抽象概念与输入方式对应,切分方式与输出模型对应;③能较好地适应不断变化的语言现象,包括结构的自组织和词语的自学习;④新知识的增加对系统处理速度影响不大,这与一般机械匹配式分词方法有很大区别[10];⑤有助于利用句法信息和语义信息来处理歧义现象,提高理解分词的效果。作为智能分词技术的一种探讨,将神经网络与专家系统思想引入中文分词,是一种有益尝试,为后续智能自动分词技术取得更多进展打下良好基础。

黄祥喜[11]提出"生成 - 测试"法,通过词典的动态化、分词知识的分布化、分词系统和句法语义系统的协同工作等手段实现词链的有效切分和汉语句子切分与理解的并行。该方法具有通用性,实现容易,分词和理解能力强。

由于汉语语言知识的笼统、复杂性,难以将各种语言信息组织成机器可直接读取的形式,因此目前基于理解的分词系统还处在试验阶段。

• 统计模型技术。苏菲[12]等提出基于规则统计模型的消歧方法和识别未登录词的词加权算法,通过词频统计、加权技术与正向逆向最大匹配进行消歧与未登录词识别。张茂元[13]等提出基于马尔可夫链的语境中文切分理论,进而提出一种语境中文分词方法,该方法建立在词法和句法基础上,从语境角度分析歧义字段,提高分词准确率。

1.1.3 *统计分词方法* 统计方法思想基础是:词是稳定的汉字的组合,在上下文中汉字与汉字相邻共现的概率能够较好地反映成词的可信度。因此对语料中相邻共现的汉字的组合频度进行统计,计算他们的统计信息并作为分词的依据。常用统计量有如词频、互信息、t - 测试差,相关分词模型有最大概率分词模型、最大熵分词模型、N - Gram 元分词模型、有向图模型等。孙茂松等提出了一种利用句内相邻字之间的互信息及 t - 测试差这两个统计量解决汉语自动分词中交集型歧义切分字段的方法[14],并进一步提出将两者线性叠加的新的统计量 md,并引入"峰"和"谷"的概念,设计了一种无词表的自动分词算法[15]。王思力等[16]提出一种利用双字耦合度和 t - 测试差解决中文分词中交叉歧义的方法。孙晓、黄德根[17]提出基于最长次长匹配的方法建立汉语切分路径有向图,将汉语自动分词转换为在有向图中选择正确的切分路径。

三种主流方法各有优缺点,其具体比较见表 1[18]。

1.1.4 *组合方法* 单个方法有优点,但也存在不足,实际分词算法设计时需要组合几种方法,利用各自优点,克服不足,以更好解决分词难题。

• 字典与统计组合。翟凤文等[19]提出了一种字典与统计相结合的分词方法,首先利用字典分词方法

表1 三种分词方法比较[18]

比较指标/分词	词典分词	理解分词	统计分词
歧义识别	差	强	强
新词发现	差	强	强
需要词典	需要	不需要	不需要
需要语料库	不需要	不需要	需要
需要规则库	不需要	需要	不需要
算法复杂性	容易	难	一般
技术成熟度	成熟	不成熟	成熟
实施难度	容易	很难	一般
分词准确性	一般	准确	较准
分词速度	快	慢	一般
应用广泛性	广泛	一般	广泛

进行第一步处理,然后利用统计方法处理第一步所产生的歧义问题和未登录词问题。该算法通过改进字典的存储结构,提高了字典匹配的速度;通过统计和规则相结合提高交集型歧义切分的准确率,并且一定条件下解决了语境中高频未登录词问题。

• 分词与词性标注组合。词性标注是指对库内语篇中所有的单词根据其语法作用加注词性标记。将分词和词类标注结合起来,利用丰富的词类信息对分词决策提供帮助,并且在标注过程中又反过来对分词结果进行检验、调整,从而极大地提高切分的准确率。白拴虎[20]将自动分词和基于隐马尔可夫链的词性自动标注技术结合起来,利用人工标注语料库中提取出的词性二元统计规律来消解切分歧义。佟晓筠等[21]设计N-最短路径自动分词和词性自动标注一体化处理的模型,在分词阶段召回N个最佳结果作为候选集,最终的结果会在未登录词识别和词性标注之后,从这N个最有潜力的候选结果中选优得到。姜涛等[22]对Kit提出基于实例的中文分词-词性标注模型,通过理论上定性分析和实验证明得出如下优点:①对于训练语料相关的文本(即与训练语料相同、相似或同领域的文本),EBST系统的分词-词性标注结果具有极高的准确率;②EBST系统的分词-词性标注结果与训练语料中的分词-词性标注具有很好的一致性。

1.2 歧义消除研究

1.2.1 歧义类型 歧义是指同一个字符串存在不止一种切分形式。歧义字段分为交集型歧义字段(交叉歧义)、组合型歧义字段(覆盖歧义)两种。据统计交叉歧义字段占到了总歧义字段的86%,所以解决交叉歧义字段是分词要解决的重点与难点。

1.2.2 消歧方法 目前解决歧义消除的典型方法有:

• 穷举法。找出待分析字串所有可能的词,该方法简单,但时间开销大,实用性不强。多数时候采用双向匹配算法,正向匹配结果与逆向匹配结果一致,分词正确,否则分词有歧义。

• 联想-回溯法。李国臣等[23]提出联想-回溯法,先将待切分的汉字符号串序列依特征词词库分割为若干子串,每个子串或为词或为词群(几个词组合而成的线性序列),然后利用实词词库和规则库再将词群细分为词。分词时,利用了一定语法知识。联想和回溯机制同时作用于分割和细分两个阶段,旨在有效解决歧义组合结构的切分问题。

• 词性标注。白拴虎[20]利用马尔可夫链的词性标注技术结合分词算法消解切分歧义,其他学者也有类似成果出现。

• EM(Expectation Maximization)法。王伟等[24]提出基于EM思想,每个句子所对应的所有(或一定范围内)的分词结果构成训练集,通过这个训练集和初始的语言模型可以估计出一个新的语言模型,最终的语言模型通过多次迭代而得到。EM是极大似然原则下的建模方法,存在过度拟合问题。

• 短语匹配与语义规则法。姚继伟、赵东范[25]在短语结构文法的基础上,提出一种基于局部单一短语匹配和语义规则相结合的消歧方法。通过增加短语间的右嵌套规则和采用有限自动机的实现方式,解决了短语规则中存在冗余项的问题,提高了短语匹配效率和歧义消除类型的针对性。

1.3 未登录词研究

1.3.1 未登录词类型 未登录词大致包含两大类:①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、团体和其他企事业单位)等[26]。未登录词识别指正确识别未在词典中出现的词,未登录词出现极大影响了分词的精度,如何解决未登录词识别问题成为分词准确性的一大难题。

1.3.2 未登录词识别 识别第一类未登录词一般是先根据某种算法自动生成一张候选词表(无监督的机器学习策略),再人工筛选出其中的新词并补充到词表中。该方法需要大规模语料库支持。第二种常用办法是:首先依据从各类专有名词库中总结出的统计知识(如姓氏用字及其频度)和人工归纳出的专有名词的某些结构规则,在输入句子中猜测可能成为专有名词的汉字串并给出其置信度,之后利用对该类专有名词有标识意义的紧邻上下文信息如称谓,以及全局统计量和局部统计量参见下文,进行进一步鉴定[26]。

归纳起来,未登录词解决方案有两大类:专用方法与通用方法。专用方法主要针对特定领域的未登录词

如中文人名、中文地名、中文机构名等识别，此类方法主要基于专有词库与规则展开。通用方法则重在解决所有类别的未登录词识别问题，前面列举的机械分词、理解分词、统计分词方法就是一种通用方法。

• 专有名词库。对中文人名、地名、机构名等分别建立词库，该方法需要搜集特定资源并制定特定算法，信息集成难度大。

• 启发式规则。通过前后缀的修饰词发现人名等未登录词。如“先生张三”，前面“先生”就是一个特定的修饰词，一般后面紧接着是人名。郑家恒[27]将中文姓氏用字进行归类，并利用分类信息建立规则以识别“小张”、“老李”之类的人名，并且有效地区分出“张”“李”等字的量词用法。

• 通用解决方案。不针对特定的未登录词设计算法，适用于各种类型的未登录词。前述三种主流分词及组合算法则属于通用解决方案。另外，吕雅娟等[28]对中国人名、中国地名、外国译名进行整体识别为目标，采用分解处理策略降低了整体处理难度，并使用动态规划方法实现了最佳路径的搜索，较好地解决了未登录词之间的冲突问题。秦文、苑春法[29]提出了决策树的未登录词识别方法，适用各种未登录词识别。

1.4 分词与词性标注评测

各种算法优劣需要在真实文本上以较大规模、客观、定量的方式进行公开公正评测，它是推动中文信息处理研究的重要手段。杨尔弘[30]等介绍了2003年“863中文与接口技术”汉语自动分词与词性标注一体化评测内容、评测方法、测试试题的选择与产生、测试指标以及测试结果，各种测试结果以精确率、召回率、F值度量，并对参评系统的切分和标注错误进行了总结。

2 分词系统研究

中文分词系统是利用计算机对中文文本进行词语自动识别的系统。一个高效的、性能优良的中文分词系统应该具备几个基本要素：分词精度、分词速度、系统可维护性、通用性、适应性[6,10]。基于分词系统特点，将分词系统研究分为早期自动分词系统与现代分词系统研究两部分。

2.1 早期自动分词系统

20世纪80年代初有学者开始研究自动分词系统，陆续有一些实用性系统出现，典型的有：CDWS分词系统[31]、汉语自动分词系统－NEWS[32]、书面汉语自动分词专家系统[6]等。由于受硬件条件及分词技术影响，早期分词实用系统在分词速度与精度上还不够理想，实用性不高。但这些实用分词系统的出现为后续分词系统设计打下了良好基础。

2.2 现代分词系统

2.2.1 中国科学院计算所汉语词法分析系统ICTCLAS ICTCLAS（Institute of Computing Technology，Chinese Lexical Analysis System）是中国科学院计算技术研究所研制，主要功能包括中文分词，词性标注，命名实体识别，新词识别；支持用户词典，繁体中文，GBK、UTF－8、UTF－7、UNICODE等多种编码格式。目前ICTCLAS3.0分词速度单机为996KB/s，分词精度为98.45%，API不超过200KB，各种词典数据压缩后不到3M[33]。

2.2.2 海量智能分词研究版 海量智能分词系统较好地解决了分词领域中的两大技术难题：歧义切分和新词的识别，分词准确率达到99.6%，分词效率为2 000万字/分钟。其中组合歧义的处理一直是分词领域的难点中的难点，海量分词系统能对绝大多数的组合歧义进行正确的切分。在新词的识别上，针对不同类型采用不同识别算法，其中包括对人名、音译词、机构团体名称、数量词等新词的识别，其准确率比较高[34]。

由于计算机硬件技术的大幅提升，分词技术的逐步成熟，现在分词系统在歧义消除、未登录词识别方面取得较大进展，分词速度与精度明显提高，实用性越来越强，为中文信息处理带来极大方便。

3 结 语

歧义消除与未登录词识别还是目前中文分词研究领域难点问题，各种算法围绕两大难题展开。论文归纳出中文分词研究热点：①创新算法，研究者需在更广泛的方法论上探讨算法，创新提出一揽子方案，设计出通用的解决歧义与未登录词识别的方法，提高分词精度与速度。②统计组合算法，目前大量文献集中于统计分词研究，基于统计的分词及与其他方法的组合是以后研究热点，将会给中文分词技术带来实质性突破。

参考文献：

[1] 黄昌宁，赵海. 中文分词十年回顾. 中文信息学报，2007，21(3)：8－19.

[2] 苗夺谦，卫志华. 中文文本信息处理的原理与应用. 北京：清华大学出版社，2007：20.

[3] 孙茂松，左正平，黄昌宁. 汉语自动分词词典机制的实验研究.

中文信息学报,1999,14(1):1-6.
[4] 姚兴山. 基于Hash算法的中文分词研究. 现代图书情报技术,2008(3):78-81.
[5] 陈桂林,王永成等. 一种改进的快速分词算法. 计算机研究与发展,2000,37(4):418-424.
[6] 何克抗,徐辉,孙波. 书面汉语自动分词专家系统设计原理. 中文信息学报,1992,5(2):1-14.
[7] 王彩荣. 汉语自动分词专家系统的设计与实现. 微处理机,2004(3):56-57.
[8] 林亚平,严锋. 汉语自动分词中的神经网络技术研究. 湖南大学学报,1997,24(6):95-101.
[9] 尹锋. 基于神经网络的汉语自动分词系统的设计与分析. 情报学报,1998,17(1):41-50.
[10] 龚汉明,周长胜. 汉语分词技术综述. 北京机械工业学院学报,2004,19(3):53-55.
[11] 黄祥喜. 面汉语自动分词的"生成一侧试"方法. 中文信息学报,1989,3(4):43-49.
[12] 苏菲,王丹力,戴国忠. 基于标记的规则统计模型与未登录词识别算法. 计算机工程与应用,2004(15):43-45.
[13] 张茂元,卢正鼎,邹春燕. 一种基于语境的中文分词方法研究. 小型微型计算机系统,2006,26(1):130-133.
[14] 孙茂松,黄昌宁,邹嘉彦,等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展,1997,34(5):332-339.
[15] 孙茂松,肖明,邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词. 计算机学报, 2004, 27(6):736-742.
[16] 王思力,王斌. 基于双字耦合度的中文分词交叉歧义处理方法. 中文信息学报,2007,21(5):15-17.
[17] 孙晓,黄德根. 基于动态规划的最小代价路径汉语自动分词. 小型微型计算机系统,2006,27(3):516-519.
[18] 未知. 三种中文分词算法优劣比较. [2010-07-28]. http://hi.baidu.com/lewutian/blog/item/b28e27fd08c0a01a08244d6c.html.
[19] 翟凤文,赫枫龄,左万利. 字典与统计相结合的中文分词方法. 小型微型计算机系统,2006,27(9):1766-1771.
[20] 白拴虎. 汉语词切分及词性标注一体化方法. 计算语言学进展与应用. 北京:北京清华大学出版社,1995:56-61.
[21] 佟晓筠,宋国龙,刘强,等. 中文分词及词性标注一体化模型研究. 计算机科学,2007,34(9):174-175.
[22] 姜涛,姚天顺,张俐. 基于实例的中文分词-词性标注方法的应用研究. 小型微型计算机系统,2007,28(11):2090-2093.
[23] 李国臣,刘开瑛,张永奎. 汉语自动分词及歧义组合结构的处理. 中文信息学报,1988,2(3):27-32.
[24] 王伟,钟义信,孙建,等. 一种基于EM非监督训练的自组织分词歧义解决方案. 中文信息学报,2001,15(2):38-44.
[25] 姚继伟,赵东范. 基于短语匹配的中文分词消歧方法. 吉林大学学报(理学版),2010,48(3):427-432.
[26] 孙茂松,邹嘉彦. 汉语自动分词研究评述. 当代语言学,2001,3(1):22-32.
[27] 郑家恒,刘开瑛. 自动分词系统中姓氏人名处理策略探讨//陈力为. 计算语言学研究与应用. 北京:北京语言学院出版社,1993:89-95.
[28] 吕雅娟,赵铁军,杨沐昀,等. 基于分解与动态规划策略的汉语未登录词识别. 中文信息学报,2001,15(1):28-33.
[29] 秦文,苑春法. 基于决策树的汉语未登录词识别. 中文信息学报,2004,18(1):14-19.
[30] 杨尔弘. 汉语自动分词和词性标注评测. 中文信息学报,2005,20(1):44-49.
[31] 梁南元. 书面汉语自动分词系统-CDWS. 中文信息学报,1987,1(2):44-52.
[32] 张永奎,李国臣. 新闻语料自动分词系统. 山西大学学报,1993,16(3):280-284.
[33] ICTCLAS汉语分词系统. [2010-07-10]. http://ictclas.org/.
[34] 海量信息. [2010-07-10]. http://www.hylanda.com/.

〔作者简介〕奉国和,男,1971年生,副教授,博士,发表论文30余篇。
郑 伟,男,1978年生,讲师,硕士,发表论文5篇。

下 期 要 目