中文标题自动抽词的递归校正法

王　维

提要　中文标题自动抽词标引是一个正在探索中的课题，本文在研究前人所取得的经验、成果的基础上，提出了一种新的实现方法：递归校正法。文中论述了这种方法的基本思想及其依据，并在初步试验中获得了预期效果。

一、引　言

文献数据库是情报检索系统的基础，标引是建立文献数据库的中心环节，在文献的存储与检索中，标引处于最突出的地位。

标引的特点是工作量大、技术性强、情况复杂，其原因在于：(1)文献数量庞大；(2)文献内容广泛；(3)用户需求多样。由此造成了存储与检索间比重的不平衡。改进标引方法，提高标引效率与质量，是进一步发展检索系统的主要研究课题之一。

从广义上讲，无论是手工标引，还是自动标引，均可归入两类：赋词标引与抽词标引。赋词标引是指从文献外选索引词，索引词既可与文献用语相同，也可不同；既可是受控词，也可是非控词。抽词标引是指从文献内选索引词，索引词必须是文献用语（关键词）。赋词标引实质上是抽词标引的延伸。相同的是，两者都须先浏览文献，确定其中的有效词；不同点是，前者再将有效词转换成系统的受控词，后者直接用有效词作索引词。

显然，为了节省人力、缩短时间，降低成本，采取抽词标引方式容易实现自动化。并且，抽词标引一般不受标引员主观意志的影响，而在相当大程度上，依赖于文献作者，它基本上忠实于文献内容，忠实于作者意图，称得上比较真实、可靠、一致的客观描述。在抽词标引中，既没有标引员的倾向与误解，也没有人为的伪装与修饰，在某种程度上，保持文献的原貌，这是抽词标引的一大优点。

实事上，采取计算机抽词标引方式，可以标引文献，也可以作文摘。因此，二十多年来，国外许多有志于此的人在不断地思考、探索与试验，力图将人的智力标引减到最小限度，甚至完全消除。

在我国，虽然对抽词标引的最一般形式 KWIC（题内关键词）索引和 KWOC（题外关键词）索引已经很熟悉了，但却很少见到论述我国如何应用这些技术的文章。至于其它形式的自动抽词标引，对国内的影响就更小了。据调查，截止于一九八一年底，尚未见到专门研究汉语抽词标引的公开文章发表，近几年来发表的有关文章也十分有限。因此，研究汉语抽词标引是我国的一项新课题。

抽词标引影响面最广的是 KWIC 索引和 KWOC索引。KWIC 索引和 KWOC 索引一般从文献题目抽词。题目在很大程度上可以反映论文的实质性内容，特别是科技论文的题目越来越表现出描述主题内容的趋势，这说明 KWIC 索引具有广泛应用的价值。

目前，有关中文标题自动抽词标引的文章发表较少，据知主要有陈培久：《篇名自动标

引的试验》，王永成：《中文标题自动抽词编制轮排索引》，王知津对抽词法的研究，杨则正一般理论研究，等等。

陈培久同志的篇名自动标引法在中情所 TK-70机上进行了试验（未见论文抽取方法的介绍），用两部机器词典实现初分词。特殊符号词典采取最长匹配法查找。汉语随机词典收词而不收片语、词组、句子，用最小匹配法查找，切分出来的词通过40个组词加工构成标引关键词。试验结果，平均标引深度为1.6，作者没有给出人工标引与自动标引所选取的标引词一致程度的数据式分析。有人认为，由于汉语语法的不规则性，篇名法通过词性判别及组合模式形成关键词标引，产生歧义的现象是难免的。作者尚未提供这方面试验的结果。

王永成同志的中文标题自动抽词标引法认为汉语词汇一般都可以拆分成一字单词与二字单词，因此他建立了一字词典与二字词典，存入每个词与其它所有可能的组合性质，如可否作词首、词尾或词中间部分，可否作独立词等等，往往一个词有多种可能的组合特性。在对中文标题进行自动分词时，根据读入的一字词或二字词的组合性质判断哪里应联词，哪里应切分。由于词典中存入的组词特性只是从语法上分析的统计结果，并没有具体指出某字词可与哪些字词相连，而且每个被收入的字词都有多种组词特性，所以在实际切分词时会出现多种可能的连接、切分方法，其中有正确的，也有错误的，因为很可能某些字词在语法上可以相连，但实际并不存在这样的词，在这种情况下，机器难于自动进行判断、取舍、在实用时似还是一个值得探讨的问题。

汉字是方块字，不是拼音文字，实现中文信息处理比拼音文字困难得多。抽词标引一般以单词为基本抽取单元，汉语词汇自动识别的问题不解决，汉语自动抽词标引便无从谈起。所以解决汉语自动分词问题是实现汉语自动抽词标引的一个关键。本文试图在这个方面进行一些探讨，并提出一种方法：递归校正法，愿与大家商讨。

二、递归校正法的基本思想

本文在设计计算机自动分词方法时，先考察了人工标引的过程和特点。人工标引时为什么能对标题中的汉字串进行分词，并区别出词的虚实呢？这是因为人通过学习，掌握了一定的知识，于是在脑中便建立起一个汉语词汇库，并且知道哪些词是实词，哪些是虚词。标引时，人们将标题中的相邻字联起来，与大脑词库中的词进行对照。对照相符，便可确定其为一个词，并随之知道它是实词还是虚词。由于汉语词汇丰富多彩，在联字、对照词的时候还须顾及前后文的关系，以防错联。常常出现这样的情况：人们读标题短语时顺次从前向后取词，当读到某一字时会发现其前某一个词联错了，往往是前面多联了一个或数个字而形成另一个词，但不符合原文意思，且使其后的字串词义不通，不好继续联词了，于是就要返回前面校正。如“联合国际力量”一短语，起初我们会将“联合国”三个字联成一个词，但继续往下读时发现不通了，于是再回头取“联合”一词。实际上人们在分析、联词时，并不是一个字一个字地读，而是一段段地看，先通读一下上下文字串，而后再决定如何将字串切分、联词，这个过程很快，在脑中一闪而过，往往被人们忽视，但只要仔细品味一下各人读文章句子时的意念，便可明白。本文提出一个计算机模拟人脑进行自动抽词的方法：递归校正法。

递归校正法的基本思想是：建立一个字典式的有效词库、一个字典式的无效词库，收集在某一专业方面的论文标题中可能出现的所有词汇。抽取题中关键词的过程为：依次取标题

中的字作为词首，先在有效词库中寻找以此字为词首的词汇，与题中词首字以后的字串对比，将题中以此字为词首的所有联词可能都暂记下来。再取最长匹配后的字作为另一词首，依上法联词。结果某词首在有效词库中未寻到，那么就去寻找无效词库，如在无效词库中寻到了依然按上法联词，并注明为无效词；如果未寻到，其原因可能有二：(一)前面的最长匹配错了，使得后面一词的词首选错；(二)词库收集的词汇不全。系统先按第一种情况处理，即取消前一词首的最长匹配，以其次长匹配为准，取其后一个字为下一词首，继续联词。这样的回归校正可以循环进行，直到联词工作进行到标题末尾，最后取各有效词首的最长匹配为标引词。如果某一次回归校正一直返回到标题首字，则说明词库收集的词汇不全，此标题的抽词工作在现有词库下不能进行下去，这时系统自动将此标题作为一疑难问题记入疑问数据库，最后在人机对话系统下完成截词工作，且系统自动将原来未收集的新词汇记入数据库。

三、关于建立词库

为什么要建词库呢？上面已分析了人脑对标题短语的分词过程，人是凭借已有的知识在脑中形成的词库与实际字串进行对比而完成分词工作的。那么在应用计算机系统进行自动分词时是否需要建立这样一个词库呢？我们知道，计算机在处理一些实际问题时可以模拟人脑，即实现人工智能；计算机也可另辟途径，以其它方法、从其它角度完成同样一个实际问题。那么，在进行中文标题自动分词时计算机应该走那条途径呢？计算机能否以某些逻辑、规则为依据进行自动分词。自然语言是人类社会中的一种交流思想的工具，它是在长期的劳动生活中逐渐形成的一种能表达人类思想的符号系统，是一种约定俗成的东西。它与数学语言不同，数字语言中存在许多严密的逻辑规则，元素之间可以通过这些规则相联系，违背了这些规则便不适应这样的数字语言系统；没有这些规则，那么数学语言系统便不复存在了。而自然语言的存在完全不依赖于语法、词法规则的存在与否，它是一种表达思想的、大家都习惯并予以承认的符号系统，也就是说，它主要是以习俗、经验为标准的。而那些语法、词法规则只是人们在研究自然语言时找出的一些适应某些语言部分的规律。自然语言只是在一定程度，某些方面体现了这些规律性的东西。但自然语言仍按其固有的形式发展，在社会生活实践中不断地增添新的语言元素，这些元素是人们在生活中约定俗成的符号。自然语言的存在与发展完全不必顾及其是否适应那些语言、词法规则。丰富多彩的自然语言中可以有许多方面不符合语法规则，但这并不影响它的存在。所以如果试图用一些语词规则来理解、处理自然语言信息，是不大可能成功的，因为这种方法不符合自然语言的本质，在丰富多彩的汉语中，那些现有的语词规则是相对贫乏、苍白的，它们不可能反映出所有的语言现象。如一个字可以出现在许多词汇中，可以出现在词首、词尾、词的中间，也可以独立成词，相邻的字符中的任何一个都可能有多种联词规则，怎样给它们联词呢？并不是每一种符号联词法则的字串都可以联成词，因为现实生活中很可能并不存在这样的词。而且词法规则并不能解释所有的语词现象，有些词虽违反构词法则，但它却是实际存在的。因此，自然语言的丰富多彩、相对的无规律性使计算机很难寻找到一般的法则来处理它。

所以，在计算机现有智能条件下，处理此问题的较妥善的方法还是建立词库。目前可供选择的也只有这种切分方法。问题在于，使用什么样的机器词典，是主题词典还是普通汉语词典。我们不否认在大容量高速计算机上实现普通词典切分词的可能性及有效性，可是作为

机器词典的兰本，能处理一切输入语句的普通词典还不存在。考虑到以能投入实用为前提，我们可以以现有的某一专业的主题词表为基础，适当加以改进，建立起适用于某一专业中文标题自动抽词标引的词库。这是一个关键词词库，它收入主题词表中所有的规范词汇及非规范词汇，并根据人工标引积累的实践经验，对可能出现于题目中而主题词表没有收进的关键词汇进行归纳、统计，一并收入机器词典中。无效词典的建立则主要是靠人工标引中的积累，统计。无效词典的通用性较强，一般不受专业的影响。

词库不可能一次就建立得完美无缺，可以设计一个人机对话系统，对实际应用中计算机遇到的难题进行人工干预处理，同时对原有的词库输入新的内容，以求不断完善。

本文设计的递归校正法需建立两个词库：有效词库和无效词库，结构一致，库中的词按字典式顺序排列，每一个词条各有三个项目：词的长度、检索匹配时的出口、检索不匹配时的出口。建立出口时以首字相同的词为一组，在组内再按各词的字典式排列划分成有限个树形结构的等级。如 a 是一级，ab，ac，ad是二级，ab下的abc，abd 是三级，ac下的ace,acf是三级等。按级次由小到大及同一级次的排列先后的顺序检索匹配。若寻找到某一级次的某一词是匹配的，再寻找这个词下属的高一级匹配，直至找到最长的匹配。

由于每一个库中词汇都有匹配成功时或不成功时的具体出口，这样，检索时就可以依据这些出口进行有针对性的寻找，避免了盲目性，大大减少了比较次数。实际上，这两种出口起的是指针的作用。本文设计了一个建库程序，词库中的每个项目都可以通过计算机自动完成。

四、关于最长匹配

语言是有相关性的，语句中相邻词语之间必然存在一定的联系。一般在不至于产生多义性误解的情况下，语句中相邻的字词联起来只要能构成一个词，我们就认为它们应联成一个词，而不将其拆开，当所考虑的词前后的字串不能与其再联成新的词时，这就是此词首字的最长匹配了，我们将这个最长匹配作为一个词抽出，这就是最长匹配法的基本思想，本文在抽词时依据的也是这个法则。例如：“计算机”三个相邻字符出现在某一句中时，“计算”是一个词，“计算机”也是一个词，但我们取其最长匹配“计算机”，将它作为一个词抽出，这样的情况下最长匹配的抽词法毫无疑义是合理、正确的。再如，在一个语句中出现“生物化学”四个相邻字符，按照最长匹配的法则，我们认为这是一个词，是一门边缘学科的名称，而不认为是“生物”与“化学”的意思。对于这种情况，只要作者采取一定措施不使读者产生多义性的误解，最长匹配法则也是可行的。如上例，若作者指的是“生物”与“化学”的意思，那么他就应在句中“生物”、“化学”两词之间加上一个连接词或分隔符（如顿号、逗号），这样，最长匹配法则就不会误解作者的意思，或者，在句中“生物”一词可与其前的字符构成词，由于最长匹配是从前向后逐步进行的，所以自然会将“生物”一词与它前面的字相连，而不会误将“生物”与“化学”相连，这时即使作者不在“生物”与“化学”两词间加连接词或分隔符，最长匹配法也不会出错。这里，我们排除需经过上下文语境分析才可得出分词结论的多义性词组。

五、关于递归校正

如上分析，人在阅读语句进行分词时，总是先从前向后扫描上下文，而后再确定分词，

在扫描过程中，人脑可能会短暂地出现一个错误的分词法，但随着扫描的进程即可很快纠正，这正是本文设计的递归校正法的依据。

计算机模拟人工，从左至右逐个扫描句中字符，先辨别出所有可能的分词法，这些词中包括最长匹配和非最长匹配，若扫描进展顺利，最后我们便可确定所有最长匹配就是从句中分出的一个个词。若扫描过程中发现前面的最长匹配有误，计算机便立即返回校正，废除原最长匹配代之以次长匹配，然后再继续扫描下去。这样的校正可以进行多次，直到不出现错误为止。通过实践中的统计分析，需要递归校正的情况并不是很多的，计算机在处理大部分语句时都很顺利，不需要返回纠正，所以这种算法的效率是比较高的。只在出现错误时才进行有针对性的局部的校正，这十分符合常理，可以省去许多不必要的机械的过程。

六、 关于人机对话系统

计算机初次建立的词库不可能十分完善，总会有一些词被遗漏，对于某些语句的分词工作不能顺利完成。本文考虑了一个人机对话子系统，计算机可以将其不能处理的标题短语暂记在疑问数据库中，最后通过人机对话子系统由人工干预完成分词工作，同时计算机自动将原词库未收入的新词收进库中，以后遇到同样的问题便可自动处理了。这样，通过人机对话系统，计算机可以不断充实原有的词库，这是一个类似人脑的学习过程，人要不断学习新知识，计算机也同样如此，通过学习，才可能趋向完善。

七、 程序框图、操作方法与试验结果

程序框图见42页。

操作方法：启动计算机后，进入DBASE状态，用USE TITLE命令打开中文标题数据库，接着输入所有待标引的中文标题。这时，准备工作完毕。当打入DO CKW命令后，系统便进入运行，这时屏幕上会自动显示出“正在运行”的字样。当所有标题的抽词标引工作结束后，抽出的各篇文献题目的关键词已全部记入答案数据库中，可以随时被调用，如用来编制索引等。如果计算机在运行中遇到了不能解决的问题，这时它便会向人提出请求帮助，在人工干预下解决这些疑难问题。至此，本批中文标题的自动抽词工作便告结束。

试验情况：作者在词库中输入了少量的词汇，对以下三个中文标题进行了抽词试验：

（1）计算机与标引

（2）联合国的概况

（3）联合国际力量

结果，第一个中文标题被抽出“计算机”、“标引”两个词；第二个抽出“联合国”、“概况”；第三抽出“联合”、“国际”、“力量”。与预期结果相符。

八、 结 束 语

本文设计的递归校正法是一种计算机模拟人工的分词方法，它的原理及整个算法过程与人脑的抽词标引过程基本一致。可以处理除可能产生误解的或须经过上下文语境分析才可确

定的多义性词组以外的所有标题短语，不会产生语法错误，也不会抽出实际不存在的词。对于不能解决的问题，可自动记下，请求人的干预。作者认为这是一个比较合理、稳妥、简捷的方法。实现这种算法时，最费力的步骤建词库，由于汉字输入符问题还没有彻底解决，所以建库时比较费时费力，但目前市场上已出现了不少种类的实用汉字词库，可以选择一种合适的现有词库，在它的基础上加以改进，而后投入使用，这不失为一种方便可行的办法。

目前的所有中文标题自动抽词标引系统都不是尽善尽美的，这还是一个正在探索中的课题。但人们可以通过多次的努力、实践，不断提高水平，终能利用计算机解决中文标题自动抽词标引问题。（本文写作过程中承蒙南京大学图书馆系邵品洪教授的指导，在此鸣谢！）