书面汉语自动分词系统—CDWS

梁南元

北京航空学院计算机系

【摘要】 本文在大量统计的基础上，论证了计算机自动分词是可行的。CDWS The Mordern Printed Chinese Distinguishing Word System) 是作者设计的一个有较高切分精度、可实用的现代书面汉语自动分词系统，它采用了词尾字构词检错技术及若干有效的纠错知识，配置了知识库和临时词典，显著的降低了错误切分率。

一、序　论

在大多数拼音文字中，词是由传统确定的，词就是字，字就是词，一般来说不存在分词问题。例如“铁路”，英语“railway”被认为是一个词，俄语“Железная Дорога”被认为是两个词，法语“chemin de fer”被认为是三个词，这并没有什么能自圆其说的道理好讲。汉语是一种没有明显的形态界限可以作为分词依据的表意语言，没有词儿连写的习惯，所以汉语存在分词问题。

汉语言的理解、汉外翻译、词频统计等自然语言处理系统都以词作为基本处理单位。因为汉语的计算机输入一般都以字为单位，因此，把输入的汉语字的序列切分为词的序列是这些自然语言处理系统必须进行的一步工作。

人工分词在精度上得不到保证。这是因为人们在阅读时，大脑有一个中间转换过程，即把视觉的形象转变成声音的形象[1] [2]。在这个过程中，存在着模糊的分词处理，它是与视觉到声音的转换和语义理解交叉或同时进行的，并以语感的形式体现出来。不过语感因人而异，由于文化水平不同，特别是古汉语修养不同，对于词和非词、词和词组的语感就会很不同，因而同一性得不到保证。1982年，北京航空学院计算机科学与工程系曾经做过一次试验，三十余个具有高中毕业文化水平的青年对五百字的一个语言材料人工分词，同一率只有50%左右。经过短期培训后，同一率也只能达到80%左右。大数据量处理时，人工分词不仅速度很慢，长时间枯燥单调的工作也使错误切分次数大大增加。这些都表明人工分词不能满足汉语言处理现代化的要求。所以研究计算机自动分词有其广泛的实际意义。

汉语自动分词分为书面汉语分词和口语分词两个方面。这里书面汉语指编码击键或自动识别输入计算机的汉字正文序列，口语指以音频信号输入计算机的汉字正文序列，二者的研究有所不同。本文只讨论书面汉语的自动分词。

二、自动分词的可行性

汉语是有层次的集合，研究汉语往往从两个方向进行：1.组合，由较低的层次到较高的

层次，如偏旁部首组合成字，字组合成词等；2.分解，由较高的层次到较低的层次，如句子分解为主、谓、宾语等。迄今为止提出的自动分词方法有五种[3][4]。宏观的看汉语分词属于分解一类的研究，从这些分词方法的具体实现看，与其说是分词，不如说是汉字试探组词。汉字构词具有极大的灵活性和自由性。由于汉语缺乏严格意义的形态变化，所以只要词汇意义和语言习惯允许，就能组合起来，没有任何限制。[3][4]所述各种分词方法，既不进行语法分析，也不进行语义理解，只机械的匹配比较，必然要出现错误切分。但是这些分词方法究竟错误切分的比率是多少，是否可以满足实际需要；错误切分的主要类型有那些；怎样才能提高正确切分率，以前研究甚少。自动分词是否可行，评价一种分词方法的精度如何，只列举出一些例子是远远不够的。这是因为人们一般认为自动分词出现的问题并不一定是自动分词中出现的主要问题。必须进行大量的统计，才能回答上面提出的问题。

为方便讨论起见，本文给出如下定义：定义一：$\alpha = a_1 a_2 \cdots a_m$ 是汉字串，其中 $a_1, a_2, \cdots, a_m$ 是汉字或字符，m 称为 α 的长度，用 $|\alpha|$ 表示，即 $|\alpha| = m$。

定义二：句子中出现的有限个汉语词的序列称为字段。

定义三：在自动分词过程中，如果一个字段有不同划分词的形式，就称出现了一次多义切分，这个字段称为多义切分字段。当一个多义切分字段被错误切分时，我们称它为错误切分字段。

只有多义切分字段才有可能发生错误切分。

定义四：设字串 β 是词 $x_1 x_2 \cdots x_i \beta$ 的后缀，又是词 $\beta y_1 y_2 \cdots y_j$ 的前缀，字段 $x_1 x_2 \cdots x_i \beta y_1 y_2 \cdots y_j$ 在汉语句子中出现过，我们称 β 为交集字串，$x_1 x_2 \cdots x_i \beta y_1 y_2 \cdots y_j$ 为交集型多义切分字段，简称交集字段。由此而产生的错误切分称为交集型错误切分。

定义五：一个交集型多义切分字段可以有一个交集字串链，交集字串的个数称为链长。

例如，字段“结合成分子时”（“—”表示词的划分）有四个交集字串“合”、“成”、“分”、“子”，这个字段的链长为四。

定义六：设 $\beta, \alpha_1, \alpha_2$ 都是词，$\beta = \alpha_1 \alpha_2$，在汉语句子中，$\beta$ 即有“$\alpha_1 \alpha_2$”形式的切分，也有“$\alpha_1 \alpha_2$”形式的切分，我们称 β 为多义组合型多义切分字段，简称多义组合字段，因此而产生的错误切分称为多义组合型错误切分。

例如，“起身”是一个多义组合型字段。在句子“他站起身来”中，它应当切分为“起身”，在句子“明天起身去北京”中，它应当切分为“起身”。

因为各种机械的自动分词方法都“取大不取小”，所以当多义组合字段应当切分为多个词时，机械的自动分词方法都产生错误切分。

交集型多义切分字段和多义组合型多义切分字段组成多义切分字段。

定义七：一种分词方法如果平均 n 个字发生一次错误切分，我们称 $1/n$ 是这种分词方法的错误切分率，简称错分率。

我们对一个22731字的自然科学样本 S_1 和一个25361的社会科学样本 S_2 进行了内容广泛的统计。下面表一是多义切分字段分布表，表二是最大匹配法—MM 方法（The Maximum Matchig Method）和逆向的最大匹配法—RMM 方法（The Reverse Maximum Matching Method）两种分词方法切分 S_1 和 S_2 的一些统计数据。

从表二可以看出，Rmm 方法精度较高，错分率约1/245。这是在不进行任何语法分析和语义理解，只机械匹配的情况下取得的，它已经能满足一些标准不高的应用要求，如果增加

样　　本	S1	S2	SALL
1. 字　　数	22731	25361	48092
2. 词　　数	14272	16540	30812
3. 多义切分字段数	325	361	686
4. 多音字引起的多义切分字段数	21	58	79
5. 交集字段数	287	231	518
6. 固定分法类交集字段数	253	210	463
7. 多种分法类交集字段数	34	21	55
8. 涉及2字以上词的交集字段数	36	6	42
9. 链长为1的交集字段数	187	183	370
10. 链长为2的交集字段数	95	48	143
11. 链长为3的交集字段数	4	0	4
12. 链长为4的交集字段数	1	0	1
13. 链长为5的交集字段数	0	0	0
14. 多义组合字段数	17	24	41
15. 字段“个人”出现次数	0	48	48

表一　多义切分字段分布表

注：5—14项都不包括4。

切分方法 / 统计项	RMM(ORM)方法			MM(OM)方法		
	S1	S2	SALL	S1	S2	SALL
错误切分次数	95	101	196	138	146	248
交集切分字段错切次数	81	82	163	124	127	251
多义组合字段错切次数	14	19	33	14	19	33
错误切分率	1/239	1/251	1/245	1/166	1/174	1/169

表二　各种分词方法错误切分分布表

注：各统计数字不包括多音字引起的交集切分字段数，多义组合字段错误切分次数不包括“个人”引起的错误切分次数。

二字人名出现次数	三字人名出现次数	二字地名出现次数	三字地名出现次数	总计字数
323	294	47	42	1748

表三　人、地名统计表

一些语法分析处理，分词精度将进一步提高。因此我们可以肯定地说，汉语自动分词是可行的。

从这次统计中，我们得到了一些重要的结论。除特别指出外，以下引用的数据均来自SALL的统计结果。

1. 对句子进行语法分析就可以正确切分90%以上的多义切分字段，其它多义切分字段也只需要对所在句子进行语义理解就可以正确切分。对于人们进行分词研究时经常提到的“乒乓球拍卖完了”等需要理解上下文语义，才能正确切分的句子，其出现概率非常低，在研究中并不占据重要的地位。

2. 多音字往往是常用字，构词能力一般较强，有12%的多义切分字段是由多音字引起的，所以在分词时采用以音为主的编码方案有较好的效果。本文介绍的现代书面汉语分词系统—CDWS (The Modern Printed Chinese Distinguishing Word System) 使用纯拼音的“词字混合码”编码方案，同形不同音的字有不同的编码，多音字为交集字串引起的多义字段不会发生错误切分，所以本文不讨论这种类型的多义字段。

3. SALL的交集字串长都为1，即：设x,y,z是汉字串，xyz是一个交集字段，|y| = 1。

4. 交集字段中涉及到二字以上的词很少，仅为全部交集字段的6.45%。

5. 链长为1的交集字段最多，是全部交集字段的71.34%；链长为2的交集字段占全部交集字段的27.61%；其它只有0.97%。没有链长大于4的交集字段。

6. 交集字段的出现次数与被处理材料的句子长度有关。样本S1平均每个句子有10.05个字，样本S2平均每个句子有7.91个字。前者平均79.20个字有一个交集字段，后者平均109.79个字出现一个交集字段。句子平均字数越多，交集字段越多。

7. 样本S2是社会科学文学类样本，其人、地名字数占总字数的6.89%，人、地名词数占总词数的4.27%。因为绝大多数专有名词不可能收入分词词典，因此在现阶段对专有名词人工进行前置处理是必需的。例如在文学类样本中处理人、地名，在植物学类样本中处理各种植物名（全世界植物近40万种）等。专有名词在汉语言材料中容易识别，前置处理工作不给操作员增加较大负担。

8. 由于样本的背景干扰，字段“<数词>个人”在样本S2中出现48次，这种情况不代表一般性。

9. RMM方法的切分精度较高，其错误切分次数是MM方法的69.01%。MM方法较不精确的原因有：

a. 形如xyz的词组都不收入词典，其略语形式为xy时却往往要收入词典。例如“高速度”不收入词典，“高速”却收入词典，于是“高速度”被MM方法错误切分为“高速度”。样本SALL中，有56次错误切分是由此产生的。

b. 单字方位词的频度较高，构词能力很强。单字方位词构词一般是在词的第一字位置，有一小部分在词的第二字位置，极少在词的最后字位置的。设x是单字方位词，字段xyz经常被MM方法错误切分为xyz；而字段yzx极少被错误切分为yzx的。在样本SALL中，有20次错误切分是由此产生的。

以上结论对我们设计CDWS有重要的意义。

三、CDWS 的实现

CDWS 是基于可靠、可实用和结构化的原则设计的一个现代书面汉语分词系统，它是在1983年实现的[5]，也是我国第一个分词系统。CDWS 采用MM分词方法，词尾字构词检错技术和知识纠错。它有两种工作方式：批处理方式和终端对话方式。终端对话方式用于处理小数据量和高精度的语言材料，检查知识以及分词词典收词情况等；批处理方式用于处理大数据量的语言材料，它又有知识库自动分词和人工干预分词两种操作方式，前者分词速度为11—15字/秒（在 HP—3000计算机系统上），对 SALL 的错误切分率为1/625，后者错误切分率约为1/1173。考虑到编码击键输入或光电输入的错误率高于1/500字，对于大多数应用来说，CDWS 是可以满足要求的。

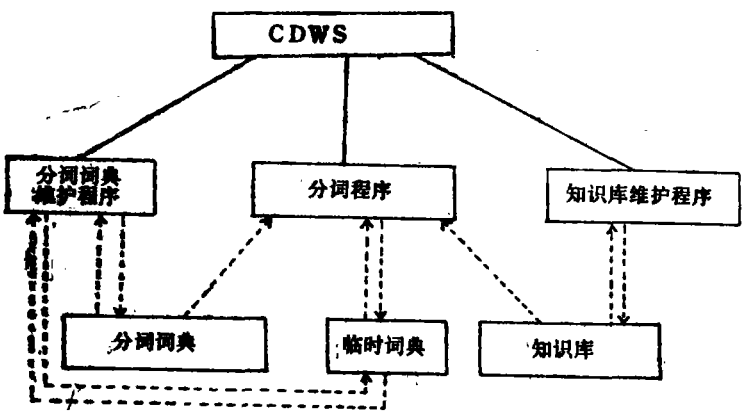
CDWS 还有如下统计功能：

1. 统计句子、标点符号数；
2. 统计字、词总数，1—7字词的总数；
3. 统计链长1—7的交集字段出现个数，知识切分数；
4. 统计多义组合字段出现个数，知识切分数；
5. 错码个数。

CDWS 是一个逻辑上独立的系统，可以用于汉外翻译、汉语言理解、词频统计等基于分词结果的语言处理系统。CDWS 只须增加六条语句，就可以同时统计字频；只须修改一个子程序，就可以同时统计词频。

CDWS 的系统结构

CDWS 的系统结构图见图一。



图一 CDWS 系统结构图

分词词典维护程序用于维护分词词典和临时词典。知识库维护程序用于维护知识库，它具有增加、删除、修改，打印知识的功能。分词程序用于分词。这三个程序实际上是相互独立的。

分词词典收词原则

国内外的汉语词典不仅收录汉语言学定义下的词，也收录非词。如果有必要收入词典的，可以不考虑其是不是词。事实上，历来的词典就是这样办的。CDWS的分词词典使用北京航空学院“词频统计”的词典也是这样收录词条的。它收录有《现代汉语词典》、《现代汉语词表》、《词海》、《汉语拼音词汇》等23本词典、辞书的词条，经过去重和删除整理，计有词条124,500，它是国内迄今见到的最大的机器词典。它除收有一般公认的汉语词汇外，还收有：

1. 专有名词，如有影响的人、地名等；

2. 简称/略语，如“政委”等；

3. 成语，如“东山再起”、“胸有成竹”等；

4. 惯用语/成语（格式与成语相似，来源和修辞色彩与成语不同），如“不三不四”，“乱七八糟”等；

5. 熟语/惯用语：

a. 动宾结构，如“拖后腿”，“打棍子”等；

b. 主谓结构，如“头疼”、“手长”等。这些词条很多是多义组合字段；

c. 偏正结构，如“小先生”等；

歇后语，如“泥菩萨过江——自身难保”，“丈二和尚——摸不着头脑”等；许多可以成为句子的警文和谚语，如：“不入虎穴，焉得虎子”，“三个臭皮匠，胜过诸葛亮”等；字数大于7的词条都没有收入分词词典。所以，分词词典中最大字数的词条为7字。

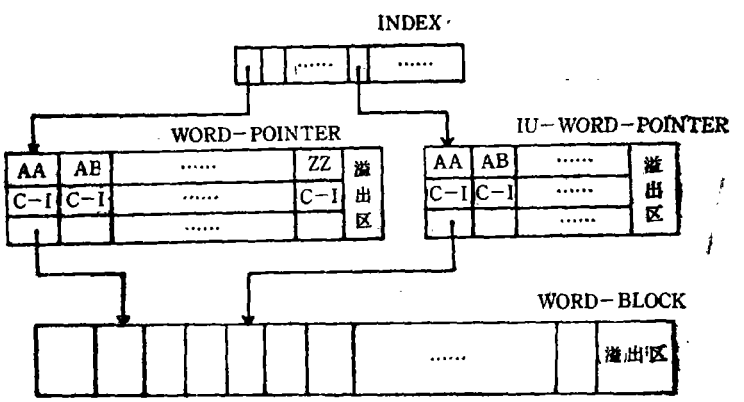
分词词典收录了一些新生词汇，如“五讲四美”、“立交桥”、“优生”等。

单字词与非词字的划分是汉语言学没有解决的一个问题。CDWS的分词词典收录的单字词依据于《现代汉语词表》的单字词表。

分词词典的数据结构

任何一种算法都有一种或数种与之相应的数据结构。CDWS的分词词典的数据结构称为词首字索引式，是为MM方法和词首字构词检错[5]设计的。

词首索引式分词词典是在文件系统上实现的（见图二）。它有分词速度较快，节省存贮



图二 词首字索引式分词词典数据结构

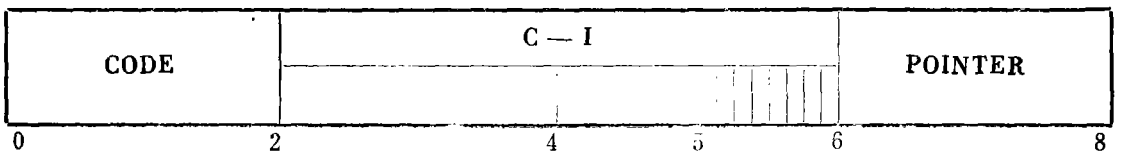
空间，易移植等优点。缺点是分词程序较复杂。

CDWS的内部码为词字混合码，一个汉字码长4位或5位，前两位都是英文字母，其余是阿拉伯数字或英文字母。5位码以“I”或“U”打头。

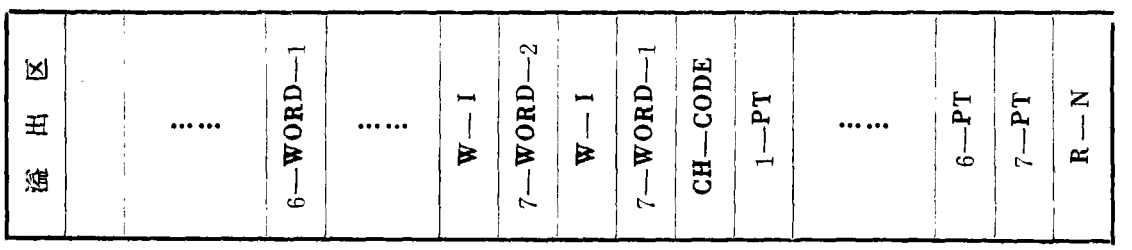
词首字索引是一个有8969个字的常驻内存的二级索引。INDEX 是676字长的一级索引，用HASH函数寻址：

$$H(LETTER1,LETTER2) = (LETTER1-65) * 26 + (LETTER2-64)。$$

其中 LETTER1 和 LETTER2 是汉字编码的1、2字符。为了节约存贮空间，二级索引有WORD—POINTER和IU—WORD—POINTER 两个部分，分别充当4码和5码长汉字的索引。WORD—POINTER的存贮结构(见图三)。其中C—I 是字项信息，长4字节，1、2字节待用，第3字节为临时词库指针，第4字节各位表示：



图三　WORD—POINTER存贮结构



图四　WORD—BLOCK逻辑记录数据结构

7. 是否单字词；
6. 是否单字数词；
5. 是否单字量词；
4. 是否单字方位词；
3. 是否前缀语素；
2. 是否后缀语素；
1. 是否切分标志；
0. 是否单字动词；

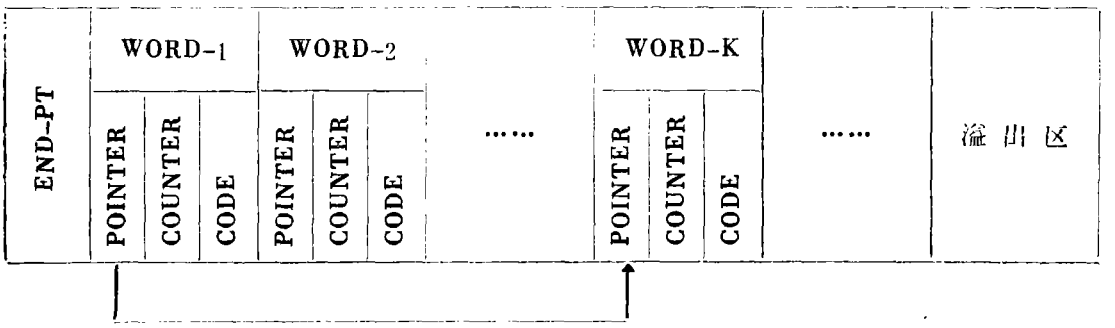
CODE 是汉字编码的第三、四字符。IU—WORD—POINTER 的 CODE 是汉字编码的第三、四、五字符。POINTER是长两个字节的指针。词首字索引共74832个字节(不包括溢出区)。

WORD-BLOCK 是一个有6100个逻辑记录的随机文件，其数据结构(见图四)。所有词首字相同的词条在同一个WORD-BLOCK 逻辑记录中。其中R-N是相对记录号；i-PT，i = 1，2，…，7 分别表示 i 字词在逻辑记录中的始址；CH-CODE 是字的编码，表示本逻辑记录都是以此字打头的词条；i-WORD-j，i = 1，2，…，7 表示第 j 个 i 字词第二字到第 i 字的编码；W-I是词项信息区和知识库指针。

WORD-BLOCK共2,356,480字节（不包括溢出区)。

临时词典

临时词典用于存贮分词词典没有的词，以减少因分词词典收词不足而引起的错误切分。临时词典以词的第一字寻址，其数据结构(见图五)。其中END-PT是末尾记录指针；COUNTER是频度计数器。



图五　临时词典数据结构

可以定期的检查临时词典各词的动态频度，把频度高的或者重要的词通过分词词典维护程序转贮到分词词典中，其余的词可以删除或继续保存在临时词典中。

分词方法

CDWS采用了结合切分标志的MM方法。在词首字索引式分词词典的支持下，它的时间复杂度是较低的。

结合切分标志的MM方法减小了每一轮匹配的初始长度，从而减少了对WORD-BLOCK的访问次数以及比较次数。经过对SALL分词时间的统计，结合切分标志的MM方法的分词速度是MM方法的1.1倍。

检错方法和纠错知识

MM方法的错误切分率为1/169。为了提高分词精度，需要检查出错误切分字段，加以正确切分。CDWS使用词尾字构词检错技术，连续构词检错7次，能够检查出链长不大于7的交集字段。据SALL统计，它检查出了占全部多义切分字段90%以上的全部交集字段。

CDWS实现了一些可以正确切分多义字段的知识（将另文介绍），它们可以正确切分SALL中75%的多义字段，因知识不足而无法切分的多义字段分两种可选择的操作方式切分：

1. 人机干预。CRT将显示出多义切分字段所在句子，操作员选择正确切分形式后CDWS继续执行。这种操作方式错分率小于1/1173。

2. MM切分形式优先策略。无法切分的多义字段采取MM方法的切分形式。这时错分率约1/625。

CDWS记录每一个多义字段的切分形式，供专家分析做参考。

随着知识库知识的增加，错误切分率将逐步降低。

五、结束语

CDWS 1983年年底交付使用，它是我国实现的第一个汉语自动分词系统，它的实现从事实上证明了自动分词是可行的。随着知识库中知识的增加，错误切分率将逐渐降低（代价是降低分词速度）。

自然语言处理的首要问题是识别语言。而自然语言的结构在实质上是一个变化无常的体系，语义上往往存在二义性。识别语言还需要计算机科学家、心理学家、语言学家等做进一步的努力。因为存在着知识表示、语法分析、语义形式化、词的划分、词类的划分等计算机科学和汉语言学领域的问题，无错误的自动分词系统近期内不可能实现。

CDWS 是以实用为目的，而非单为实验设计的一个自动分词系统。从对七百余万字的分词处理情况看，它达到了这个目的。CDWS 的实现将对汉语言的理解、汉——外翻译、词频统计等的研究起较大的作用。

本文是在刘源副教授的指导和帮助下完成的，在此致以衷心的感谢！

参考文献

[1] Ovid J. L. TEeng, "Speeching Recording in Reading Chinese Character", Journal of Exprimental Psychology, 1977。

[2] Simon, H.A., "信息的存贮系统——记忆"。

[3] 梁南元，刘源，"书面汉语的计算机自动分词"，中文信息，1986年第一期。

[4] 梁南元，刘源，"OM自动分词方法"。中文信息，1985年第3期。

[5] 梁南元，"汉语的自动分词与一个自动分词系统——CDWS"，全国汉字处理系统学术会议，1983年12月。

[6] 王以德，"程序辅助分词初探"，中文信息研究会基础理论专业委员会第一次学术会议，1982年。

[7] 管纪文等，"结合上下文辅助分词的学习系统"，中文信息研究会第二届学术会议，1983年5月。