语言和计算机 3

语言和计算机

《语言和计算机》编辑组 编

3

LANGUAGE AND COMPUTER

中国社会科学出版社

语言和计算机

（3）

《语言和计算机》编辑组

中国社会科学出版社

责任编辑：黄　华
责任校对：王　新
版式设计：韩　锐

语言和计算机
YUYAN HE JISUANJI
(3)

中国社会科学出版社出版
新华书店北京发行所发行
八九九二〇部队印刷厂印刷

787×1092毫米　16开本　11.75印张284千字
1986年7月第1版　1986年7月第1次印刷
印数 1—4,200册
统一书号：9190·032　定价：2.05元

目　录

Main Articles

论机器翻译规则系统的编制方法

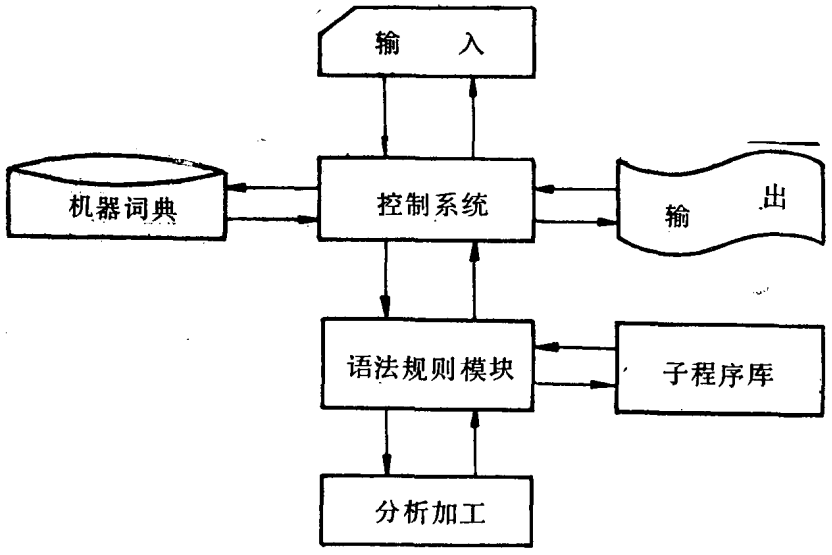
刘　倬

1. 前　言

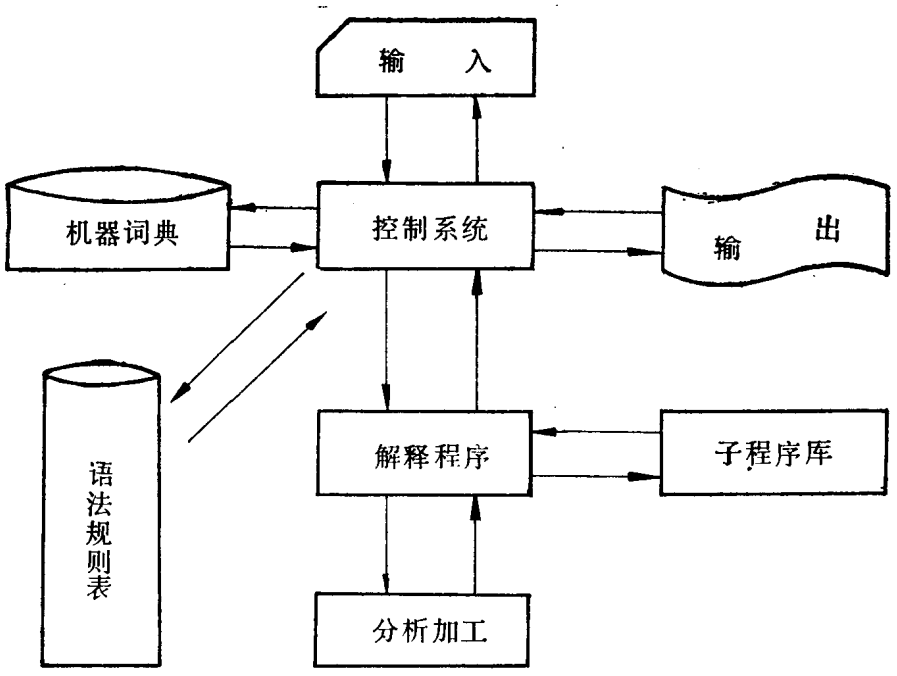
机器翻译包括语言分析，特别是源语分析和转换分析，以及程序设计两个方面。其中，语言分析是机器翻译成败的关键，这是人们公认的一种看法。但是，绝不能由此得出结论说，程序设计只是处于无关紧要的消极从属地位。因为，语言研究提出的分析、转换和综合规则是机器翻译规则系统的内容，而程序设计提出的算法是这些规则在计算机上运算时的表达形式。语言分析研究和程序设计之间的关系，实质上就是内容与形式的关系。在规则的内容和规则的表达形式这一统一体中，内容决定着形式，而形式也明显地影响着内容。研究和编制机器翻译规则系统，如果只强调前者而忽略后者，势必给研制工作带来不良的后果。程序设计是否合理，算法是否优化对于规则的精确性、抽象性和实用性，以至于对整个机器翻译的实现都有着直接的影响和作用。

按程序设计的算法不同，机器翻译规则系统可分为两大类：二元翻译规则系统和三元翻译规则系统。

二元翻译系统工作原理：



三元翻译系统工作原理：



所谓二元翻译规则系统，指的是系统由两个元素——机器词典和机器语法构成。过去我们编制的60型俄汉机器翻译系统，JFY-I型英汉机器翻译系统，JFY-Ⅱ型英汉机器翻译系统和现在正在编制的JFY-Ⅲ型英汉机器翻译系统都是二元规则系统。所谓三元翻译规则系统，指的是系统由三个元素——机器词典、语法规则表和解释程序构成。采用这种算法时，不同翻译系统的区别只是机器词典和语法规则表的内容不同，解释程序可以作为一个通用程序，用于任何系统中。1981年我们曾就这种方法提出了一套语法规则表和解释程序的设计，并进行了程序的调试和试验。今后，待条件允许时还准备在JFY-Ⅲ型系统中作进一步的试验。

机器翻译程序设计的算法，一方面和翻译规则的内容有关，另一方面和程序设计时使用什么语言也有联系。本文主要是就用COBOL程序设计语言编制机器翻译规则系统遇到的一些问题提出讨论，不妥之处，希大家指正。

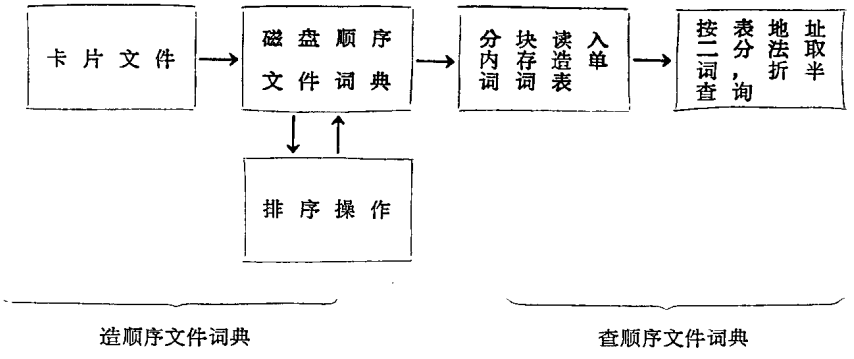
2. 机器词典的编制方法

2.1 文件的形式和综合词典的造查方法

一般说，COBOL程序设计语言有四种不同的文件：顺序文件、相关文件、索引文件和直接文件。文件的种类不同，造查词典的方法也有明显的区别。下面以综合词典为例，分别介绍一下各类文件词典的编制方法。

2.1.1 顺序文件词典的编制方法

对这种词典来说，理想的编制方法之一是顺序造-折半查的算法。这种算法的程序简单易行，适合于进行小型或中型机器翻译试验。顺序造-折半查算法的工作流程可图示如下：



这里，物理读的块长视计算机内存的大小而定，一块包含的记录越多，读取的速度就越快。但内存容量小时，要注意给单词词表留够存储空间。为了减少内外存载体转换的次数，提高查词典的速度，在折半查的基础上还可以引入一种“批处理”的技术。所谓“批处理”，包含三个含义。其一，顺序文件词典中的词可分段分批地调入内存造表。其二，查词典时，原文句子中的词也可分段分批地查词典。如何分段分批，可以根据与表尾词（这在造表时用特定的数据项给出）比较，随机确定。比表尾词小的词，是当前表中应该有的词，属于这一批要查的当前加工词，比表尾词大的词可用暂时放过的办法，留作以后批处理的查询对象。其三，当前加工词在词典中查到后，还可继续查询该词至句末或调入的原文段末有无与该词相同的词，如果有，就把在词典中取到的该词的所有信息传送给找到的相同词。为了提高查准率，这种查询可以仅仅用于虚词和某些常用词的加工上。采用上述“批处理”的办法，可以保证磁盘顺序文件词典读入内存一遍，就查完一个句子的所有词。如果使用的计算机内存容量比较大，查词典时还可以把一次调入一个句子扩充到一次调入一批句子，这一批句子是五个、十个，还是更多个，可根据设备的容量大小而定。

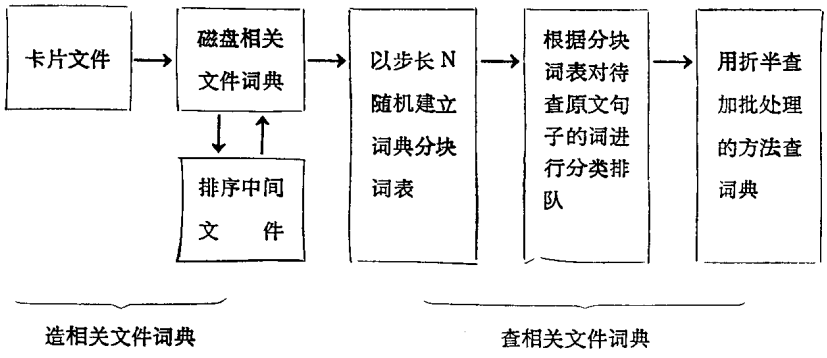
另外，在折半查的算法中必须给出表长的函数。这可以用不定长和定长两种形式给出。所谓不定长，就是在确定表尾词时，把表尾词的地址下标值赋给表长函数。所谓定长，指的是内存中定义的表的长度就是表长的函数。如果在最后一次造表过程中，记录装不满表的所有空间（这是经常发生的），就用“假记录”填满表的剩余部分。“假记录”就是给记录的所有字节都置高位值（HIGH-VALUE）。如果，在表初始化时不用零或空而用高位值，那么，在造表过程中装不满记录的所有表空间就会自然而然地变成假记录项了。有了假记录项，一张没有填满记录的表也会是一个记录项值从小到大的顺序文件。这样一来，用定长折半法查词就不会出现差错了。

顺序文件词典的优点是，造查词典的程序简单，容易调试，增补新记录也不困难；缺点是，词典修改起来比较麻烦，特别是不能直接删除某条已经存入的记录。必须删除时，只有用替换的办法，即用一条新记录（或假记录）替换待删除的记录。

2.1.2 相关文件词典的编制方法

COBOL程序设计语言中规定相关文件有两种存取方式：顺序存取和随机存取(有的文本中还规定有动态存取)。采用随机存取方式建立相关文件词典，词典可以用随机造-折半查的算法。在这种算法中，由于造词典时使用了随机方式，不仅是词典的建立，而且词典的更新、

修改、删除、增补都很简便易行。它能满足大量动态数据存取的各种要求，这是中型或大型机器翻译试验可用的一种编制词典的方法。随机造-折半查算法的工作流程可图示如下：



对相关文件词典来说，最简单易行的查词典的方法是，既不要词典分块词表，也不对待查词进行任何分类排队，而是直接用折半法，按相关地址查词典（为了使用折半法，在建立相关文件词典时不要忘记登录记录总数）。这样做唯一的缺点是内存和外存之间载体转换的次数多些，影响查词速度。为了克服这一缺点，前一节（2.1.1）提到的批处理技术在这里也完全适用，不过实现批处理的方法与前者稍有不同罢了。这里，批处理使用的方法是，根据步长N取词，给词典进行分块，这类似于通常所说的分词典。用步长为N取到的词造一个分块词的词表，这类似于分词典的地址索引表。查词典前，首先从顶到底查分块词的词表，给原文句子中的待查词进行分类排队；查词典时，按分类排队的顺序依次取词，在分块词表中找出大于该词的词，这等于该词在相关文件词典中的结束地址，结束地址减步长常数N，等于该词的开始地址。一次调出开始地址到结束地址间的所有词，用折半法查找该词以及与该词同类的词。用这种方法，为了查出原文句子中的所有词，有时需要查遍整部词典的每个块，但在大多数情况下，只需要抽查词典的某几块就够了。与顺序文件词典比较，相关文件词典能进一步提高查准率和查词速度。目前，在我们编制的ECTRAN系统中，综合词典的查词方法采用的就是类似的方法。

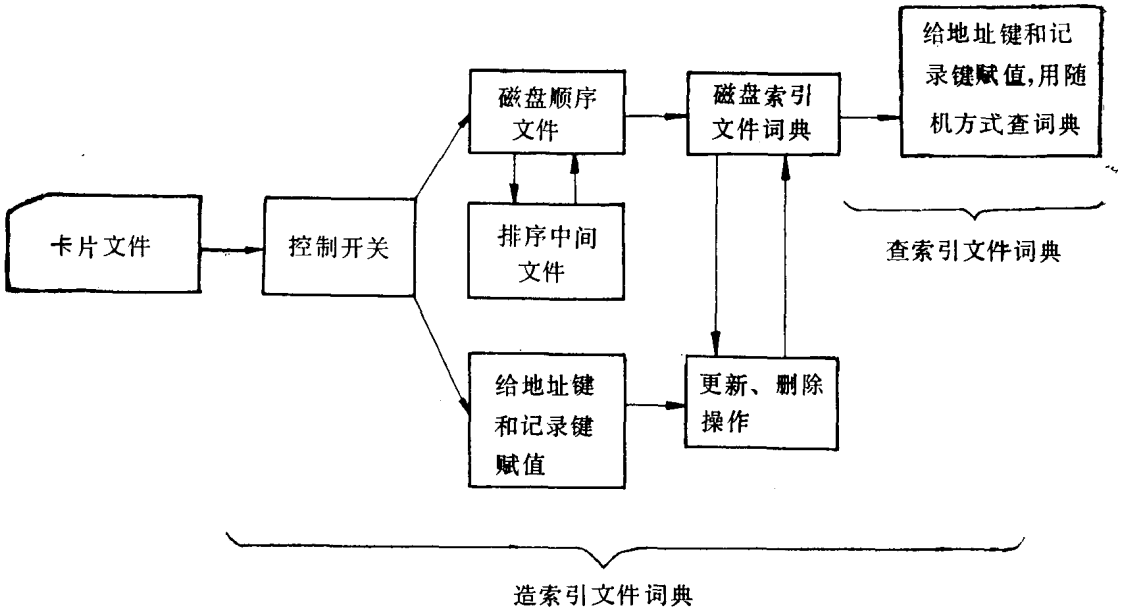
3.1.3 索引文件词典的编制方法

索引文件是COBOL程序设计语言中特有的一种存取数据的形式。① 编制索引文件词典时，系统软件会自动地向用户提供包括索引表在内的一种适合于大量数据存取需要的数据结构。这是大型或实用型机器翻译系统可用的一种编制词典的形式。如果我们把单词（词的原形）作为索引项，即记录键的值，把单词的其他所有信息作为索引项的内容，那么，就可以很容易地编制出一部理想的索引文件词典来。编制这样的词典时，把造词典的程序和更新词典的程序分开，用顺序编造-随机存取和更新的算法是比较理想的。这种算法的工作流程可用图表示(见第5页图)。

从流程图中可以看出，索引文件词典有几个优点：（1）建立词典的程序简单；（2）维护、更新方便，如果把单词作为记录键的值，那么根据卡片文件记录定义的格式写出单词及其需要的信息，给地址键赋值后就可以随机增补插入、修改或删除某项词条记录；（3）查词典的手续与前两种词典比较有明显的简化，只要把待查词的原形赋值给地址键和记录键，通过一

① 这里所说的COBOL，是指TK-70机上的BOL和FACOM-230机上的COBOL的综述。

次查词典就可以查出该词，这里不需要附加任何批处理的方法就可以保证较高的查准率和查词速度。在JEY-Ⅲ型英汉机译系统中，综合词典以及其他词典采用的都是这种编制方法。



2.1.4 直接文件词典的编制方法

直接文件需要两个参数，即两个取词键——地址键和符号键。地址键表示记录的存取区，符号键表示存的或取的是什么记录。用这种方法编制词典有极大的灵活性，只要给两个存取键赋值后，不仅可以随机直接查词典，而且还可以随机直接编造或更新词典。就随机编造词典而言，这是上述三种词典都无法做到的。用直接文件编造词典，关键在于用什么方法给两个存取键赋值。在JFY-Ⅱ型英汉翻译系统中综合词典采用的就是直接文件的形式。当时我们是用词首、词长、词的字母值的叠加和词尾合成的压缩码加上散列法的一种算法，给地址键赋值，用待查词的原形给符号键赋值。最近，与傅爱平同志合作，又在2,197个词的范围内对符号键的赋值方法做了进一步的试验，提出了如下几项试验数据，作为改进原有算法的参考。

（1）词的压缩码：首字母＋词长＋字母叠加值＋尾字母

（二位）（二位）（三位）（二位）

散列地址码：取个位数法

试验总词数：2,197

其中：两次重码418

三次重码131

四次重码22

五次重码12

六次重码 5

（2）词的压缩码：同（1）

散列地址码：首(个位)＋长(个位)＋字母叠加值(个位)＋九位压缩码叠加值

试验结果：两次重码353

三次重码68

四次重码11

五次重码 2

六次重码 0

(3)词的压缩码：同(1)

散列地址码：余除法

试验结果：两次重码327

三次重码59

四次重码 5

五次重码 0

六次重码 0

(4)词的压缩码：同(1)

散列地址码：九位数自乘取中

试验结果：两次重码331

三次重码46

四次重码 2

五次重码 0

六次重码 0

(5)词的压缩码：首＋词长＋正序第四个字母＋叠加值＋尾

(二位)(二位)　　(一位)　(三位)(二位)

散列地址码：十位数自乘取中

试验结果：两次重码330

三次重码43

四次重码 4

五次重码 0

六次重码 0

(6)词的压缩码：首＋词长＋正序第三字母＋叠加值＋尾

(二位)(二位)　　(一位)　(三位)(二位)

散列地址码：十位数自乘取中

试验结果：两次重码302

三次重码31

四次重码 4

五次重码 0

六次重码 0

这里应该指出，做试验的2,197个词中包括不少变形词。如果减去这些变形词，散列地址的重码率将会显著地降低。

2.2　成语词典、语义多义词典和功能多义词典的编制方法

2.2.1 各类词典编制方法的统一问题

在机器翻译系统中除综合词典外还有多部不同类型的词典，如成语词典（也可称作词组词典）、语义多义词典、功能多义词典（也可称作结构词典），等等。过去，在我们编制的几个系统中，不同的词典都是用不同的方法编造的。这里提出一个问题，不同的词典能否用统一的方法或接近于统一的方法编造呢？回答是肯定的。通过调查我们发现，编制不同的词典，处理的内容虽然不同，如综合词典的内容是单词和它的信息特征，成语词典是成语的形式和它的信息特征，语义多义词典是语义多义词和它的分析规则与结论，功能多义词典是功能多义词和它的结构式与结论，但是，抽去每部词典处理的特定值，余下的共同的东西就可以作为建立统一算法的依据。在不同类型的词典中余下的共同东西是什么呢？这里可以归纳出一个词典数据结构的抽象模式，即〈词形〉+〈信息特征〉。依靠这样一个模式，我们就可以把编造成语词典、语义多义词典、功能多义词典的方法与编造综合词典的方法统一起来了。在ECTRAN和JFY-Ⅲ型系统中都是这样做的。下面就以随机造-折半查的相关文件词典的编造方法为基本算法，介绍一下除综合词典之外其他各部词典是如何编造的。

2.2.2 成语词典的编制方法

成语词典可以用随机中心词存取法编造。所谓随机中心词存取法包含两个意思：其一是，对不同的成语来说，中心词在成语中的位置可以不同；其二是，查成语词典时，从中心词查起，直至查到整个成语。这时，一条成语的记录必需包括以下几个内容：中心词，中心词在成语中的位置，构成该成语的词的数目，成语词组，成语的意义及其信息特征。其中，成语词组在卡片输入文件中是按自然语言原有形式不定长表示的，而在磁盘相关文件中用三个十位字节定长的基本项表示。为了把一个不定长的成语词组转换成一个三十位字节定长的字符串，做了这样的规定：(1)成语词组的首词存入第一个基本项中；(2)如果成语词组小于或等于三个词，就把第二个词存入第二个基本项，第三个存入第三个基本项；(3)如果成语词组大于三而小于七个词，就把后两个基本项按四个字节一块分成五块，然后从成语词组的第二个词开始，每个词都取前四个字母，依次存入各个块中；(4)如果成语词组大于或等于七个词，就按先进先出的原则，用第七、第八、第九……个词从第一块起依次更新各块的内容。这样，成语词典的一条记录就可以由不定长转换成定长形式了，同时这种定长形式的成语记录还可以进一步归纳成与词典数据结构抽象模式相一致的形式：〈词形＝中心词〉+〈信息特征＝中心词的位置、词数、成语词组的定长形式、词义及其他信息特征〉。这里，用2.1节列举的造综合词典的任何一种方法，同样也能造出成语词典来，区别主要表现在查词典的方法上。查成语词典时，用折半法只能查出成语的中心词，这还不等于查到整个成语。为了查到整个成语，首先必须根据中心词提供的信息，在原文句子中找出哪几个词可能和该中心词构成成语。举例说，在原文加工区中'A''B''C'三个词是成语（如IN ADDITION TO），其中，根据综合词典提供的信息我们知道'B'是中心词。这个成语在原文加工区的安排如下所示：

原文加工区的地址1 2 3 4 5 6 7 8 9 10 11 12

原文加工区的单词△△△△A B C△△△△

⇓

成语中心词

用折半法在成语词典中可以查到中心词为'B'的这条记录。根据成语词典提供的中心词位

置，用一个公式〈中心词在原文加工区的地址－中心词位置＋1〉，求出该成语在原文加工区中的首词，这里是6－2＋1＝5。然后，根据成语词典提供的“构成成语词的词数”取词。取到的词就是可能构成成语的所有词。如果取到的一串词经过定长转换处理与词典中成语词组的形式一致，就说明原文中三个词是成语，否则就不是成语。

2.2.3 语义多义词典的编制方法

语义多义词典通常是由一组组分析规则及其结论组成的。例如，ADDITION一词，作为语义多义词有如下一组规则：

序号	规 则	是转	否转
(01)	该词为复数	05	02
(02)	该词后一词为名词	06	03
(03)	该词后一词为介词'OF'	04	05
(04)	该词后一词为材料名词	06	05
(05)	011605 '添加物'		
(06)	010503 '添加'		

同时，一个多义词的分析规则和其他多义词比较，其数目有多有少，差别很大。这种分析规则能否转换成词典数据结构的抽象模式呢？从上述例子中可以看出，一个多义词包括多少条分析规则虽然是不定的，但在计算机中每条规则都可以写成定长的，如上述规则在计算机中的存储形式可以写成：

序号	规则类型	放过	查询对象	查询类型	查询内容	是转	否转
01	#	00	G	T	090100	05	02
02	#	−1	Z	L	010000	06	03
03	#	−1	Z	Z	100101	04	05
04	#	−1	Z	S	011700	06	05
05	¥	011605		00065	000		
06	¥	010503		00063	000		

（其中：#＝分析规则，¥＝结论，−1＝后一词，
G＝该词，Z＝找到词，T＝特征，L＝类，
S＝类属，Z＝类属组）

把分析规则转换成这种定长的形式，每条规则当作一条记录，这样，多义词典数据结构的抽象模式也可定义成：〈词形＝多义词〉＋〈信息特征＝分析规则和结论〉。'ADDITION'一词的规则改造后就变成：

0010	ADDITION	#	规则
0020	ADDITION	#	规则
⌇	⌇		⌇
0060	ADDITION	¥	结论

改造后的这种规则，按照与综合词典类似的方法造词典就没有什么困难了。这里，每条记录（一条规则或一个结论）都是相对独立的，造词典时不要求人工预先排序。待所有记录都写入磁盘文件后，由程序根据词形和序号两个数据项自动排序，因为以步长为10设置了序号

项，词典的更新(增加、删除和修改)也很方便。特别是增加，只要把新的规则用词形和序号项表示出插入的位置，写入词典，经过控制分段排队，新记录就能插到预定的位置上。查词典时，首先用折半法在词典中找到当前加工的多义词，根据词形和序号检索出该词的第一条记录，然后依次取出它的所有记录，送内存工作区造表，利用解释程序进行表处理加工。应该指出，为了更有效地进行表处理加工，最近又对JFY-Ⅱ型系统中使用的多义词规则表作了一些修改。这主要表现在，用定向层次加工法替代了'是转'和'否转'的条件转移查找法。所谓定向层次加工法指的是，在规则表中增设了'层次号'一项内容。解释执行多义词分析规则时，不管查询的结果为'是'还是为'否'，总是从表的顶端向底部依次取规则执行，依次取哪条规则按层号的不同来确定。作这样的修改，不仅使多义词的分析规则用起来很方便，更重要的是使多义词分析的解释程序及其算法和语法分析规则的解释程序及其算法统一起来，做到同一个解释程序可以在不同的场合，按不同的目的，用不同方式进行调用。

2.2.4 功能多义词典的编制方法

功能多义词典是为强支配关系词和特殊用法词专门设置的一部词典。例如，APPLICATION作为功能多义词，为了查询它的强支配关系：<APPLICATION>+<OF>+<N_1>+<IN>+<N_2>，在功能多义词典中给出了这样一条查询规则：

功能多义词	支配介词	结论1	结论2	结论3
APPLICATION	IN	准谓	前介宾B	介宾C
		'应用于'	OF='把'	IN='…中'

这样一种形式的规则，同样也可以用词典数据的抽象结构模式来表达：<词形=功能多义词>+<信息特征=支配介词，结论1、2、3>。既然如此，造功能多义词典的方法，也就可以和造其他词典的方法统一起来了。查词典时使用的方法与语义多义词典的查询方法类似，也是用折半法检索出当前加工的词，把查到的词典记录调入内存后，按功能多义词的结构类型不同，用不同的子程序(或过程段)加工处理。

从以上的介绍可以看出，本文讨论的词典编制方法，与JFY-Ⅱ型系统中使用的方法比较，明显的区别在于：在JFY-Ⅱ型中使用的是地址拉链法，换句话说，不同的词典通过地址指示字链接成一个统一的词典系统。在这种词典结构中各部词典之间存在着一种链接关系，动一点就可能牵扯到全局。这给词典的维护和更新工作带来许多麻烦。本文讨论的词典编制法，依据的是词典数据抽象模式的设想，用基本相同的方法编造不同的词典。这不仅可以使造词典的方法尽可能地统一，而且每部词典，甚至其中的每条记录都是完全独立的。这给词典的维护和更新工作提供了极大的方便。

2.3 词典的存词形式问题

存词形式问题包括三方面的问题：一是存原形还是存变形，二是各类词典中存词形式要不要统一，三是定长存词还是不定长存词。在综合词典中，一般说存词的原形比较经济合理。如果在原文句子中遇到变形词，通过削尾恢复原形后再查词典。其他几部词典怎么办呢？在其他几部词典中仍旧按原形存词，这办法看来是可取的。但是，为了避免查变形词时削尾还原等繁琐的加工手续，提高查准率，把综合词典加工时查到的每个词的原形（更确切地说即词典形式)都记在原文句子的加工区中。这样一来，查其他几部词典时，不管当前加工词是原形还是变形，都可以直接按原形查词典了。特别应该指出的是，成语词组中的每个词也都用原形表示，这会给人们编写成语词条带来一些不便（这种不便是可以在程序上想办法解

决的)，但给查成语词典带来极大的方便，权衡利弊还是可行的。另外，大家知道，COBOL程序设计语言提供了定长记录和不定长记录两种存取功能，其中，还是定长存取用起来比较方便。不过一个词按定长存取，定多长是值得研究的一个问题。统计表明，英文词1—15个字符的居多，大于15个字符的是少数(虽然在科技术语中稍多一些)。确定词长时不必考虑词的完整性，而只考虑排他性，即存多少字符就可以把一个词与其他词区分开来就行了。我们认为，从这个角度出发把词长定义成12或15个字符就可以了。以这个长度存词，既可以把不同的词区分开来，也不算浪费存储空间。这里应该指出，这种有限定长的存词办法只是对磁盘文件词典而言的。在卡片文件中仍然可以按词的自然长度书写。输入时，通过词形数据项传送的自动取舍，不定长的符号串(词形)就转换成定长的符号串了。另外，在编词典时把综合词典的词长和其他各部词典的词长统一起来，其他各部词典的折半查询就显得更为方便。

2.4 词典中词的分类问题

在机器翻译中，如同在传统语言学中一样，词的分类问题是一个很棘手的问题，同时也是一个极端重要的问题。词的分类是进行任何类型的语法分析的依据和基础，它直接关系到语法分析规则的查准率和查全率。从机器翻译的实际需要出发，把语言词汇分为十类，还是二十类，还是更多的类，本文不想讨论这样一些具体问题。这里只想就机器翻译中使用的分类法的特点及其应该遵循的几个原则，提一点不成熟的看法。

(1)在机器翻译中使用的词的分类法应该是一种语法-语义分类法。语言既然是一种交流思想的工具，而且句子又是一个相对完整的思想的表达，那么，句子中句素与句素之间的语法关系必然是人们交流思想时的事物、现象和动作之间逻辑语义关系的具体表现。从这个意义上说，撇开这些具体的事物、现象和动作，只靠句子的形式结构来建立语法分析规则是不够的。经验证明，分析句子时，形式结构只能帮助我们把句子分解成大小不同的语言片段，要想进一步识别语言片段间联系的性质和意义(特别是在合成句子时)，必须分析它们在语义上的结合律、替换律和分布律。词的语法分类的语义化和语义分类的形式化，这种分类才能满足精确分析的需要，才能满足语言识别和语言合成的需要。在机器翻译中建立语言的分析、转换和综合规则也是如此，必须采取形式结构和语义结构两者兼顾的原则。为了贯彻落实这一原则，必须从功能分析的不同需要出发，根据形式和语义上的结合律、替换律和分布律的异同对词进行分类。这种分类我们就称之为语法-语义分类法。

(2)机器翻译中的词的分类法应该是一种随机动态分类法。在传统语法中词的分类往往是一种静态的，与具体上下文无关的抽象分类。这对不跨类的单义词是合适的。对同形词，特别是同形的多义词，抽象分类往往有二义性，同时，这种分类不能与用词造句的规则建立精确的、必然的联系。为了克服这一缺点，在机器翻译中应该使用一种随机动态分类法。所谓随机动态，包含两个意思：其一，词在词典中可以属于两种或两种以上不同的类或属、组，同时每一类词都有自己特有的区分规则，例如，同形词本身就包含一套区分同形的规则，语义多义词也要靠它的分析规则确定自己的词义和属、组划分；其二，在翻译过程中，词类的确定不仅仅是词典的任务，换句话说，查词典的工作和语法分析不应截然分开，一前一后进行，而应该使它们有机地结合在一起，使查词典的工作贯穿在整个语法分析中，分析到哪类词就用哪类词查词典(如果需要查词典的话)。这样可以使词的个性用法和语法分析的共性规则有机地结合起来，使它们各自的作用都得到充分的发挥。同形词或多义词的分类实际上是一个随机动态的筛选过程，它们的词类划分和语法分析同步进行比较合理，特别是功

能多义词和虚词（结构词）更是如此。这可以保证语法分析的查全率和查准率，提高翻译质量。

（3）机器翻译中词的分类法应该是一种虚实结合的分类法。在任何自然语言中都能分出两类词来。一类用作事物、现象或动作的名称，它们有明显的实物意义，传统语法中称之为实词。另一类表示事物、现象或动作之间的关系和联系，换句话说，它们只有明显的语法意义，传统语法中称之为虚词。实词在语言中几乎是无限的，分类时只能用归纳法，而虚词是有限的、可数的，分类时可以用枚举法。在机器翻译中用枚举法给虚词分类，对于语法分析来说有极其重要的意义。在分析原文句子时，只要能弄清每个虚词的用法，句子的结构分析就不会出现太大问题了。而为了弄清每个虚词的用法，进一步说，为了建立实词和虚词连用的规则，最好的办法是把各个虚词枚举出来，这样就可以很方便地归纳出一套实词和虚词连用的结构式，同时也能总结出一套虚词的分析规则。这是语言形式结构分析的重要组成部分。实词按归纳法分类，虚词（更确切地说，结构词，它不仅包括全部虚词，而且还包括在形式结构分析中有特殊作用的某些实词）按枚举法分类，我们把这种分类法称作虚实结合的分类法。

（4）机器翻译中词的分类法应该是一种相关转换分类法。任何分类都有一定的目的性。在机器翻译中词分类的目的是非常明确的，它是为句子的分析、识别、转换和合成服务的。一个类是否能成立，不仅要看它在原文分析和识别中有没有用处，而且还要考虑到，在源语到译语的转换中，甚至在译语句子的合成中它能否作为一种区别特征而起作用。例如，英、汉语处所短语的词序不同，英语中说‘北京中国’，而汉语中说‘中国北京’。为了进行这种分析和转换，名词的‘类’下可以分出一个‘属’来，表示处所名词，而处所名词又可根据它表示的处所概念的大小分成若干‘组’。这样一来，处所短语的转换分析和译语的合成就好办了。分类时从原文分析出发，又考虑到译文综合的需要。按照这种原则建立的分类法，称作相关转换分类法。

（5）机器翻译中词的分类法应该是一种横向相通的分类法。一词多类是自然语言中普遍存在的一个客观事实。这样的词只有在具体的上下文中才能确定其具体的类。在机器翻译中，词类的划分虽然比传统语法中精细得多，但对某些词，在综合词典中也无法给出唯一的类来。这里不可避免地会遇到一个词类转换分析问题。为了在这种分析中做到类的转换不影响属、组的用法，可以考虑在不同的类下设置统一的属和组。例如，‘CHEER’是个名动同形词。作动词时，在我们的分类体系中，在‘振奋’这一义项下可划为050401，其中05表示动词，04表及物，要求单宾语，01表‘精神状态’。为了适应同形分析时词类转换的需要，在名词分类中设置了010401一项。如果在具体上下文中‘CHEER’作名词用，只把它的类05改成01就可以了。在名词(01)类下的属组0401表示‘CHEER’和介词‘OF’连用(～OF＋N)，转换成汉语既可处理为动宾结构（振奋＋名），也可以翻译成偏正结构（N＋的＋振奋）。名动同形词可以这样做，名形同形词、形副同形词……等也都可照此办理。这种类与类之间属、组划分的对应和统一就是横向相通的分类原则。

（6）机器翻译中词的分类法应该是一种纵向相关的分类法。纵向指的是在一个类下属与属，组与组之间的关系。例如，在名词这一类下可以列若干个属，每个属下又可列许多组。这些属或组的排列也必须有个原则。比如说，可以按先抽象后具体的原则排列。还拿处所名词为例，组越大，它表示的地名越小。当然，不同的类，允许用不同的原则，如动词可按及

物、不及物以及动宾结构的分布特点排列，而形容词或副词可根据功能一致的原则排列，等等。属和组的有序排列，就是我们所说的纵向相关的分类原则。这样分类不仅对语法分析有用，另外给程序设计也提供了很多方便。

2.5 综合词典和其他分词典的分合问题

过去我们参加编制的几套机器翻译系统中，机器词典一般都包括综合词典、成语词典、语义多义词典以及功能多义词典等几部词典。这种划分主要出自这样的考虑，即各部词典的任务、内容和加工方法不同，当然它们的编制方法也应该不同。现在，随着机器翻译研究的深入和发展，我们开始认识到，机器词典的任务，概括地说，是提供词的形态、语义、结构，甚至转换等方面的信息特征。从语言学的角度来看，这几方面的信息是相互依存，相互影响，相互作用的一个对立统一体。生硬地把它们分开，有时是困难的，或者说区分的标准和原则是模糊的。在实际工作中我们常常为一个语言现象的处理犹豫不定，究竟按成语，还是按语义多义词，究竟算语义多义词，还是算功能多义词，这就是困难和模糊的表现。是否能把各部词典，特别是成语词典、语义多义词典和功能多义词典统一起来，合并成一部所谓"用法词典"呢？既然我们可以把各类词典的内容归纳成一个统一的'词典数据结构的抽象模式'（这解决了一个造词典方法的统一问题），那么，为什么不能从词典的查询内容中概括出一个'词典查询的抽象模式'来呢？经过一年多的探索，我们建立了一个词典查询的抽象模式，并为它作了程序设计。有了这样一个模式，各部词典的合并就不困难了。各种词典的分合问题，是值得研究的一个很有意义的问题。我们深信，随着这一问题研究的深入和发展，新的词典编造法必然向旧的编造法提出越来越严重的挑战，最后取而代之，使机器词典的编造技术提高到一个新的阶段，新的水平，从而使机器翻译的整个研究工作有一个新的突破。对机器翻译的早日应用来说，这是有战略意义的一步。我们愿和大家一起，群策群力，相互协作，为这一目标的早日实现共同奋斗。

3. 语法规则系统的编制方法

3.1 在语法规则系统中不用和少用放过规则是实现精确语法分析的必要条件之一

过去我们编制语法规则系统，多半都是自觉不自觉地从传统语法的概念和方法出发，进行成分或成分转换的'动态线性分析'，即各种放过规则控制下的上下文分析。乍看起来，这种分析方法的灵活性很大，能应付千变万化的不同情况。但是问题还有另一方面，灵活性越大，分析的精确性和准确性就越难保证。这一缺陷的典型表现就是，规则系统只管分析，不管对错。所以，有时会出现一些料想不到的问题，某个语言现象本来很简单，分析的结果会得出错误的结论，而有的语言现象看起来很复杂，但分析的结果却得出正确的结论，换句话说，在分析中会出现某些张冠李戴的错误。这是为什么呢？问题就在于'动态'上，即放过规则的使用上。一般讲，放过规则都是根据普遍存在的语言现象归纳和概括出来的。但是，语言中往往有特殊的现象，这种特殊现象是放过规则概括不了的。既然如此，利用放过规则控制上下文分析，发生张冠李戴的错误就只能看成是正常的，不可避免的了。多年的实践证明，放过规则不仅建立起来很困难，使用起来也是模棱两可，难于捉摸。把放过规则看成是语法规则系统中的一匹'害群之马'，它好作用起得不多，坏作用发挥得不少，这种估价不是言之过分的。为了建立一套精确性较高的语法规则系统，使它不仅具备分析能力，而且拥有一定