基于搭配词典的词汇语义驱动算法*

唐泓英 姚天顺

(东北大学计算机科学与技术系，沈阳 110006)

摘要 本文首先阐明了汉语分析中所要面对的问题，并描述了如何建立搭配词典以表达个性的语言现象及处理规则. 为了适应自然语言中的随机性和规律性，我们提出一个词汇语义驱动算法，它不仅提高了系统的效率，而且较好地解决了自然语言处理中诸如词汇兼类处理这样的难题.

关键词 机器翻译，汉语分析，词汇语义，词典，语法.

汉语是世界上最古老的语言之一，它自身存在着很多非常复杂的语言学问题，很难提出一套完整的形式语法规则. 本文以汉语分析为例，有意提出一种基于搭配词典的语言处理算法. 在我们汉英双向翻译系统(CETRAN)的实践中证明，该方法不仅适合汉语分析，也适合英语生成、英语分析、汉语生成. CETRAN 是一个已有一定翻译能力的实用系统，它包含1个近8万词条的汉语词典，和4万多词条的英文词典，2个专业词典，2个成语词典，4个搭配词典，及2 000条共性规则. 它在 UNIX，MS DOS 和 MS WINDOWS 3.1 环境下皆可运行，其速度是非常快的.

1 兼类与词义

任何一种语言，无论是程序语言还是自然语言都有形式与内容两个密不可分的部分. 在语言形式上，表现为语法，在语言内容上，表现为语义[1]. 如果一种语言，象程序语言，语言形式完全决定了语言的意义，那么机器对这种语言的理解是无二义的. 作为一门自然语言的现代汉语是一种口头形式和书面形式都充分发达的语言. 因其重表意而缺乏形态，仅对它进行语法分析远远不足于理解句子中各成分的逻辑语义关系，因而对汉语的语义分析是汉语分析中重要的一环. 而词汇语义的确定对整个句子的语义分析起着不可忽视的作用. 如“有”虽只有一个动词 v 的词性，在句子中只做谓语，其意义却有多个：(a)表示所属：(vv50)“人有两只手”；(b)表示存在：(vv51)“窗外有座狮子山”；(c)表示发生或出现：(vv52)“有新的情况”；(d)表示估量或比较：(vv53)“水有二米深”；(e)泛指：(vv54)“有人来晚了”；(f)用于动词前，表示客气：(vv55)“有请今晚嘉宾”. 当“有”出现于句中时，分析器要确定它在句中应是

* 本文 1994－05－16 收到，1994－07－25 定稿

作者唐泓英，女，1967 年生，助教，主要研究领域为自然语言理解，机器翻译. 姚天顺，1934 年生，教授，博士导师，主要研究领域为计算语言学，知识表示.

本文通讯联系人：唐泓英，沈阳 110006，东北大学计算机科学与技术系

哪一种语义.每句前小括号中的字符串是"有"每种语义的语义标识号,它来自基本词典.在基本词典中,对每个词的每种语义都会有一张工作单,其中有一个语义标识项,是长度为四的字符串.第一个字符表示该工作单的词性,前三个字符合起来是其下位词性,而第四个字符是对该词所有相同下位词性工作单的一种编号.举例说明,假如一个词有两种词性,名词n和动词v,名词有两种语义,相应的下位词性都是nn1,动词有三种语义,相应的下位词性是vv1,vv2,vv2.那么该词共有五张工作单,相应的语义标识号是nn10,nn11,vv10,vv20,vv21.这样,每个词形,其各种语义都有一个唯一标识其意义的代码.由这种编码方式可以看出,当一个词的语义确定下来后,其对应的词性也会确定下来.那么,该词在句子中的语法成分就会比较容易分析出来.象(a)句,如果在分析的过程中选择了语义标识号为vv50的这张工作单,则可以得知其词性是动词v.在我们的汉英双向翻译系统中采用的是句法语义一体化的分析方法,比较符合人类理解自然语言的过程.然而,如何去确定词汇的语义呢?这往往要用个性规则来完成.在(a)例中,可以由个性规则:

(1111;1113;113)+^'有',^ *(-134)⇒@setmark(vv50)

这条规则表示,当前词(^)"有"的左边是人类(1111)或动物类(1113)或组织类(113),并且其右边某处(^ *)不存在表示外形特征(134)的词,则选择语义标识为vv50的这张工作单.这便是在语言处理中常常遇到的兼类问题.如果句子中的每个词的词义确定不下来,分析就无法进行下去;即使词义确定下来,但却是错的,势必影响到整个分析的结果.所以兼类处理的质量是汉语分析成败的关键.

汉语缺乏形态变化,造成兼类困难.由于个性规则的精确性,利用个性规则是解决这类问题的一个良好途径.然而,汉语词汇量巨大,大量的词都需要兼类处理.另外,依照汉语句子结构上比较清晰的五个层次:词素、词、短语、单句和复句,我们在分析策略上,有分词、词法分析、句法分析(语义分析)及复句分析.在这几个层次的分析中,即使各词汇的兼类问题能够解决,词汇之间的语法语义关系也存在着众多特性的相互制约关系.这就提出了词语搭配问题.

2　词语搭配及搭配词典

我们知道,汉语极少有词形变化的约束,而在语义搭配问题上却显得格外突出和复杂.其实在任何语言里,词语搭配都是一个重要问题,在汉语中尤其突出.从语法上讲,语法现象:动词+名词(或代词),产生动宾关系,并非一定是正确的分析,因为不是所有动词都能与所有名词或代词搭配,也不是所有名词或代词都能与所有动词搭配[2].如果盲目地按语法类分析短语结构及句子结构,将会产生错误结果.如:

学校　请(v)　的　乐队(n)　的　水平(n)　很　高.

学校　买(v)　的　张家(n)　的　麻花(n)　很　香.

这两个句子的结构十分相似,都含有"动词(v)+的+名词(n)+的+名词(n)"的语法结构,但不同词却有不同的结合方式和修饰关系,仅凭这种语法形式很难确定动词是作用于前一个名词还是后一个名词.这种关系看起来很随机,但它反映了词汇之间的搭配关系,即"请"可以与"乐队"搭配,而不能与"水平"搭配;"买"与"麻花"而不能与"张家"搭配.这涉及到词的个性特征,不是汉语基本词典能体现的,也不是一般的编码方式所能表达的.然而,用

规则来描述并处理这些复杂细腻的关系则显得轻而易举.

词语搭配有很强的个性,在汉语中极少有搭配能力完全相同的词,所以搭配规则应该是很多的.若将所有搭配规则放入规则系统中,势必造成系统效率降低,乃至规则爆炸.为此,我们构造一个搭配词典,将这些属于个性的众多搭配规则组织起来恰当的运用.

搭配词典的总体可形式表示为

$$Col(w)=\langle cat,mor,syn,msy,sen\rangle$$

其中 cat、mor、syn、msy、sen 分别描述词的兼类,词法、句法(语义)、嵌套,句子修饰共五类处理规则集,我们称每类为词典中的区.如 cat 区、mor 区等.每个区放的是零个或零个以上同一类规则.搭配词典中的词是所有具备个性特征的词,一部分是虚词,大约有 1 500 个词,其余是实词,现已有近 2 000 词.不是每个汉语词汇都会有个性规则的,如果一个词在以上这五个处理过程中有个性表现,则在相应的区中就有一条或若干条规则来描述对它的处理.倘若一个词的所有特征和属性都在共性范围内,那么它是由共性规则来处理的,在搭配词典中就没有关于这个词的信息.搭配是广义的,不仅指语法上的形式搭配,任何两个以上可以共现的词都可以做成一条搭配规则,包括远距离的搭配.由此可见,搭配词典的词汇量不可能象基本词典那样庞大.反过来,即使每个汉语词汇都有搭配规则,也不会影响系统效率.因为在分析过程中,分析器不是去匹配所有搭配规则,而是匹配在句子中出现的那些词的搭配规则.因为在分词结束后,句子中每个词的搭配规则(如果有的话),其各区的规则以队列的形式挂到词结点上去.

搭配规则是由我们为系统设计的一种规则描述语言 CTRDL(详细书写规范请见文献[3])来描述的,与共性规则库中的规则是同一个规则描述语言.只是搭配规则常常注重词形,因词形本身蕴含着根本的语法语义信息,即使其语法语义成分没有确定下来或确定下来了但并不一定正确,那么,利用词形会更准确,也便于校正.如:

```
lex:大概
cat:^ '大概'+('的';n)⇒@setmark(a);
cat:^ '大概'+(m;p;v;a)⇒@setmark(d);
cat:q+^ '大概'⇒@setmark(n);
mor:^ '大概'+'的'|n⇒!^1,@modi(^,a),-^1,1^\^__(EXP0);
mor:^ '大概'+n⇒@modi(^,a),~^1,1^\^__(EXP0);
mor:^ '大概'+(m;p;v;a)⇒@modi(^,d),=-^1,1^\^__(MOD);
mor:q+^ '大概'⇒1^\^__(QNT);
```

这是个简单的例子,'大概'这个词是个虚词,它只有 cat 和 mor 区规则.其中,'+'表示词结点的邻接关系,'@setmark'是 CTRDL 中确定工作单的函数,它可以以词性、下位词性或语义标识作为关键字.'a','n','m','p','v','d','q'分别表示形容词,名词,数词,介词,动词,副词,量词.'@modi'表示对某一结点重新确定它的词性或词义.cat 区的规则,判明'大概'在不同句子里的词性及词义,它的每种词性,只有一种词义.在这种情况下,只需以其词性来选择工作单就可以了.

这里需要特别讲的是,由于分词的过程与其他类分析过程不同,我们主要采用自动分词的方法,当有歧义切分的可能时,也用规则分词.规则分词的原理与搭配词典规则的原理类

似,也是建立一个歧义切分词典,在自动分词的过程中,如果遇到可能有歧义切分的词,就按它的歧义切分规则去切分.与汉语分析相应,在英文生成、英文分析及汉语生成过程中都建立了各自的搭配词典,只是词典中的各区也相应地有其特别的作用.

3　词汇语义驱动算法

词汇语义这个概念,一方面,它提升具备多重属性的词汇在分析中的地位;另一方面,它强调各类众多属性之中,语义属性在语言分析过程中所起的重要作用.搭配词典的建立为词汇语义在语言处理中的驱动作用奠定了基础.

3.1　一些概念

提出词汇语义驱动算法之前,我们先介绍几个概念.

(1)结构态

结构态是被分析语言单位(词汇、短语、句子、文本等)的现有状态的总和,包括初始态、中间态及终止态.初始态是非活跃结构态;中间态是活跃状态,它有可能向另一个结构态转换;终止态亦是非活跃状态,也是一个新的初始态.结构态包括两个方面,一是语言单位中的数据,二是语言单位中的结构.句子分析中有以下几种基本结构态:

分词态:句子处于分词阶段的结构态.

兼类态:句子处于兼类处理阶段的结构态.

词法态:句子处于词法分析阶段的结构态.

句法态:句子处于句法语义分析阶段的结构态.

整句态:前四个阶段结束后句子的结构态.

(2)信息级

对于各类语法信息和语义信息在使用它们之前将其组织成类,划分成级.如图1所示:

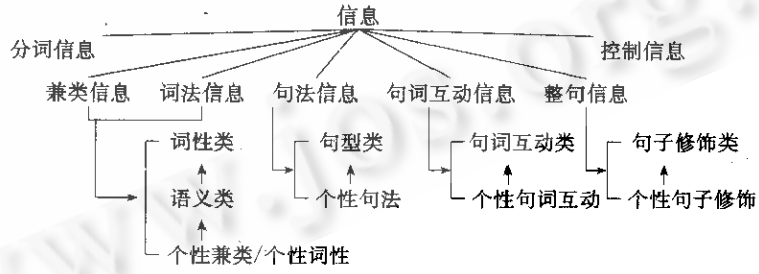


图1　信息级

(3)条件基

在信息级中的每一级,都有一系列条件和操作体系,每一种条件叫作条件基.

(4)继承性

在信息级中的两级间如有箭头关系↑,则它们具有继承关系.如果两级间有继承关系,在上一级能做的操作,在下一级中一样能够满足.比如,词性类信息集有它的的条件和操作体系,语义类与词性类有继承关系,语义类可以增添自己新的条件操作集,但在词性类中能满足的条件及操作,在语义类中一定也可以满足.

(5)摘取性

在使用信息过程中,我们有时将信息整体中一些必要的、无歧义的部分抽取出来,而忽视其内部其他属性,用抽取出来的这些无二义的信息作为根本,以确定正确的内部属性或结构态之间的关系.

在系统开发中,为了解决兼类处理问题,我们常常可以利用语言形式所表达的语法语义来使语言单位间的内部信息一致. 如,我们只需知道一个词的表象及作用,对于其具体是哪一种词性及语义可以不去确定,只可通过该词形与其他词搭配来触发正确信息. 象上面"大概"的例子,在兼类过程中如果它赖以判断的左右条件被其它词屏蔽,就有可能选择错误的工作单. 然而在词法分析的过程中,利用摘取性,只取词形本身这个绝对正确的信息作为判断的条件,并使用@modi,重新选择正确的工作单. 这种方法是对兼类处理的一个有力的补充.

3.2 词汇语义信息驱动

信息驱动算法用一个基本图可以表示为:

$$\text{结构态1} \xrightarrow{\text{信息}} \text{结构态2}$$

其中信息来源于信息级表,是指向某一级的指针,当它不空时,说明其所指规则集中某个条件基为真.

分析器不断的搜索词汇语义信息,一旦出现可使结构态发生变化的信息,就触发结构态的变化. 分析是有层次的,这几个层次并非独立的,而是相互影响相互依赖,我们称之为结构制约性. 另一方面,各层次之间也是相互嵌套的. 句法中蕴含词法现象是显而易见的,而词法层中常包含句法结构,表现为各种从句现象,我们称这种现象为结构嵌套性. 各个层次有一个最根本的分析基础,就是兼类处理,实际上,兼类过程也不是一个独立的过程,它既是几个分析层次的基础,又依赖于这几个分析层次. 所以,在整个分析策略上,就是围绕五个分析层次及兼类处理展开的.

每个层次有其自己特定的信息,当每个层次再也找不到能使其结构态发生变化的信息时,该层次就达到一个终止态;若层次间有嵌套或制约关系,那么层次之间也有信息,该信息又能触发层次内部产生新的信息,而使层次内的结构态发生变化.

图 2 是一个整句态,当它达到一个终止态时,分析器就完成了对简单句的分析. 如果要分析的句子是复杂句,则以下是对其的分析.

$$\text{整句态1} \xrightarrow{\text{局法信息}} \text{整句态1}$$

现在说明各种信息的对应关系.

由于分词采用的是自动分词方法,在自动分词过程中不需要特别的信息,但当分词的结果有歧义的可能时,分词信息便会触发与歧义段相关的规则(来自于歧义切分词典),通过匹配歧义切分规则,可以得到正确的分词结果. 另一类的分词信息来自于词法分析,因为经过一些分析后,分析器就可以确定重新合并那些本来是一个词但却在自动分词中被分开的词.

兼类信息使两个兼类态发生变化. 在变化的过程中,至少有一个词的特定词性或词义被确定下来. 兼类信息有两级. 一是词的结点上的 cat 指针,当它不空的时候,便成为兼类信息. cat 所指的规则集是从搭配词典来的个性规则. 如果第一级不存在,兼类信息的另一级就

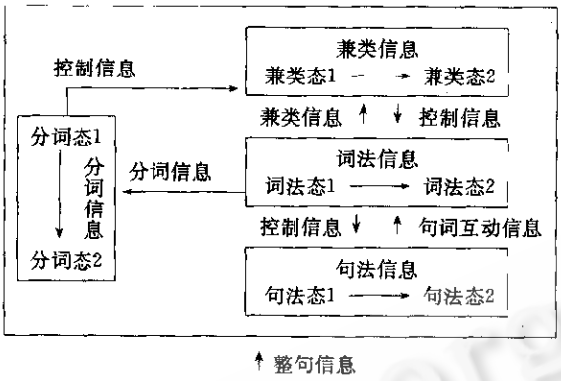


图2 简单句的词汇语义驱动图

起作用,指向词性类,并取词性类中22个兼类规则库中若干个库.类似地,词法信息、句法信息、句词互动信息、整句信息也分两级,第一级分别是结点上的mor、syn、msy、sen队列指针,另一级都是指向各类的公共规则库.无论是哪一级,每一类信息当它们不空即所指队列集合中或公共规则库中有为真的条件基时,便成为相应的可以发生作用的信息.词法信息所指的规则全部是处理词法层语言现象的.句法信息在第一级时,如果结点的词性是动词v,则在syn中描述该动词的各种句型框架及语义格关系.如果结点的词性是连词c,那么它描述该连词在复句中的作用.当结点是介词p,syn也用来描述自由介词.句法信息在第二级时,指的是各种句型库及复句处理规则库.句词互动信息是当词法中有句法嵌套时,它就被正确的条件基激活,将句法成分识别出来,并分析出它是某个词法成分的语法语义成分,这时有可能会激活新的词法信息.整句信息指向句子修饰现象的处理规则,象对'突然'、'然而'这类词的处理,单独的连词,标题结点和段落结点的处理.

当没有其它类活跃的信息存在时,控制信息便起作用,它使句子从一种结构态进入另一种结构态.

这个算法是依照自然语言本身所具有的层次结构运转的.另外,由于结构嵌套性和结构制约性,它不是机械地单向分析句子,而是每步分析着眼于句子中最可靠的信息点,以它来驱动分析,逐步使分析范围扩大,使结构态发生变化,产生更多正确可靠的信息.而最可靠的信息在于词形本身,它来源于搭配词典.所以该算法对自然语言分析有一定的通用性,至少在CETRAN中它也用于英语分析.

4　一个词汇语义驱动的例子

我们的汉英双向翻译系统CETRAN采用的是中间语言翻译方法,即将源语分析成中间语言,再经过生成而得到目标语.关于中间语言请见文献[4].这里是个汉语分析的例子.

汉语句子:"大家为她在归国的飞机中意外昏倒而担心."这个句子无分词歧义,所以利用从右向左最大匹配法即得到正确的分词结果:

大家(r)　为(p,v,v,v,v)　她(r)　在(d,p,v,v,v)　归国(v)　的(u,n)　飞机(n)　中(f,v,v,n)　意外(a,n,d)　昏倒(v)　而(c)　担心(v)　.(g)

分词后,句子中每个词汇在基本词典里的内容和在搭配词典里的个性规则全部取出挂到词结点上去.括号中的符号是与词典工作单相应的词性,如'为'有1个介词,4个动词,4

个动词的语义用英文分别表示为“regard as”,“become”,“act as”,“be”. 凡分词后有多于一张的词典工作单的词,就需对它进行兼类处理. 控制器不断查询兼类信息,‘为’,‘在’,‘的’在兼类信息的第一级确定下来,‘为’的兼类结果是介词(p),‘的’的结果是助词(u),而‘在’是副词(d),这是个错误结果,应该是介词. ‘中’和‘意外’是兼类信息的第二级确定下来的,分别为方位词(f)及副词. 之所以不同的词在不同的级被处理,仅仅在于它们的兼类规则是由共性特征决定还是需个性条件制约.

此时,句子中每个结点都有唯一的一个词性和语义,但不见得都是正确的. 这是因为,此时对句子还没有作词法句法及语义分析,兼类规则赖以判断的条件常常是这三类分析的结果,然而,这三类的分析也是建立在兼类分析的基础之上的. 这便是结构的制约性. 有了搭配词典,这类问题是可以解决的.

兼类之后,兼类信息为空,兼类态达到一个终止状态. 控制器搜索信息,得到控制信息不空,句子进入词法态的初始态. 除了‘在’,其他词都是在词法信息的第二级被处理的. ‘在’的处理是在词法信息的第一级,匹配的是‘在’mor 区的个性规则:

^ ‘在’+(n;s;f),^ 1#(v;a)⇒@cpd(^ 1#,^),^ 1\^ 1#__(LOC),! ^

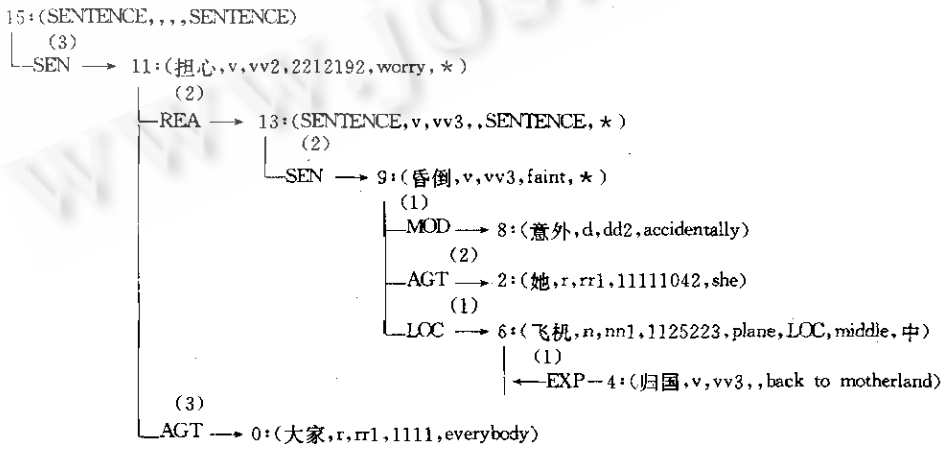
这条规则没有判断‘在’的词性,而是将其后的名词(n)或地点词(s)或方位词(f)改成后面某个动词(v)或形容词(a)的儿子,格关系为 LOC,并把‘在’删掉. @cpd 是使在此之前所得的有关‘在’的动态信息保留下来.

这时词法信息为空,词法态达到一个终止态(1)(为了说明,下面图中也标注相同数字以表明处理层次),由于句词互动信息不空,即‘为’msy 区中有条件基为真,规则为:

^ ‘为’+…+‘而’+((v;a;z),23113)⇒@gsyn(^ ‘为’,^ #‘而’),¯^ 1,1^ \^ __(REA)

这条规则使得由词法分析得到的森林的根结点“她 昏倒”做句法分析,句法信息不空,“她 昏倒”的局部句法结构态发生变化,并达到一个终止态(2).

此时控制信息又发生作用,使句子进入句法层,句法信息的第二级不空,“大家 担心”做句法分析,最终使句法态进入终止态(3),即是我们所需的中间语言:



从这个例子可以看出一个粗略的流程,但机译系统为翻译提供的各种机制并没有完全

地体现出来. 因为分析不同句子需要不同方面的信息，对信息中每级的需求也不同.

5 结　论

我们知道，当一个孩子呀呀学语时，他的知识体系是逐词建立的，当他真正能理解语言时，头脑中所反映的是语言中每个词的属性及应用规则. 机器理解语言也是一样，我们有必要使机器对句子中所有的词都能理解. 实践证明，词汇语义驱动方法在机器翻译系统中是行之有效的. 通过这种方法，我们把个性规则放入词典，从而减少了共性规则系统的负担，同时提高了整个翻译系统的效率. 系统的优势在于，它有较高的速度和精确度，在解决兼类处理问题上有特殊功效；它的弱点是，对于某些人来说，规则描述语言 CTRDL 似乎难了点儿.

参考文献

1 Tang Hongying, Yao Tianshun, Kou Yuxin. The lexical semantic driving algorithm in language processing. In: Proceedings of the 1994 International Conference on Computer Processing of Oriental Languages, Taejon, Korea, 1994. 333－338.

2 张寿康，林杏光. 现代汉语实词搭配词典. 北京：商务印书馆，1992.

3 Wang Baoku, Zhang Zhongyi, Yao Tianshun. Rule description language CTRDL in machine translation system. In: Proceedings of 1991 International Conference on Computer Processing of Chinese and Oriental Languages, Taipei, 1991. 264－269.

4 刘东立，唐泓英，姚天顺. 汉语分析器中语义网络表示方法. 中文信息学报，1992，6(4)：1－10.

THE LEXICAL SEMANTIC DRIVING ALGORITHM BASED ON THE COLLOCATION DICTIONARY

Tang Hongying　Yao Tianshun

(*Department of Computer Science and Technology, Northeastern University, Shenyang* 110006)

Abstract　This paper describes how to set up collocation dictionaries to express the individual language phenomena and processing rules, which are important complementary knowledge bases for basic dictionaries and general rule bases. In order to fit for the combination of the randomness and regularity in natural languages, the paper proposes the lexical semantic driving algorithm with an example in Chinese analyzing, which can solve the difficult problems such as polysemantic processing in natural language processing.

Key words　Machine translation, Chinese analyzing, lexical semantics, dictionary, grammar.