广义依存关系和汉语自动分析

傅 爱 平
中国社会科学院语言研究所，北京 100732
E—mail: fuap@linguistics.cass.net.cn

摘要 本文介绍一个自动分析汉语句子的实验系统。这个系统用广义依存关系表示汉语句子表层和深层的结构，采用词汇规则和语法规则相结合的方法实现层次分析、语义特征分析和语义指向分析，识别句子成分之间的结构关系和语义关系，分析的结果可以用于汉语自动理解或汉外机器翻译。
关键词 自然语言理解 汉语自动分析 依存关系 句法结构 语义关系 机器翻译

Generalized Dependency and Chinese Parsing

Fu Aiping
Institute of Linguistics, Chinese Academy of Social Sciences
E—mail: fuap@linguistics.cass.net.cn

ABSTRACT This paper presents an experimental Chinese parser based on generalized dependency relations of constituents in sentences. Having incorporated lexical and grammatical analysis, the system provides syntactic and semantic information on surface and underlying structures of Chinese simple sentences. This information could be useful to Chinese to foreign language machine translation and Chinese language understanding.
KEYWORD: Chinese parsing, Dependency relation, Syntactic structure, Semantic relation, Machine translation

1 用广义依存关系表示句子结构

依存关系的概念来自依存语法理论（冯志伟，1987；胡明扬，1988），这种理论认为，组成句子的词与词之间存在着一定的联系，所有这些联系构成了句子结构的格局。由这种结构上的联系而形成的成分之间的从属关系是句子的基本关系，一个句子的句法结构就是由其组成成分之间的各种从属关系构成的层级结构。

依存语法所揭示的语言结构规律在自然语言处理的研究中得到了广泛的应用。不少自然语言理解、机器翻译系统应用依存语法模型分析和生成语句。我们曾经在一个实用的英汉机器翻译系统中用依存关系表示英语句子的内部结构（刘倬等，1989）。在这个系统多年的翻译实践过程中，这种表示方法已经经过了大量各种类型英语语料的检验。本文介绍的汉语自动分析实验，也是在以这个系统的翻译引擎为基础，经过改造而形成的调试环境下进行的[①]。

分析和识别句子的结构是计算机理解自然语言的基础。不同的应用目标对分析的结果有不同的要求。与外汉机器翻译相比，汉外机器翻译对汉语分析的要求比较高。我们经常看到有些汉英翻译系统输出的译文与源文的意思或规范的英语相去甚远，主要原因恐怕就是源语分析还不够到位。所谓到位，在采用直接法或转换法作为翻译策略的系统中可以理解为，分析的结果有助于解释源语和目标语之间必要的对比差异。譬如，汉语的动补结构在翻译成英语时应该转换成什么句式？该句式的结构成分与汉语动补式的结构成分之间有什么关系？转换的条件是什么？例如汉语句子“昨天来的那个孩子哭红了眼睛。”，我们的实验系统把它的结构表示为：

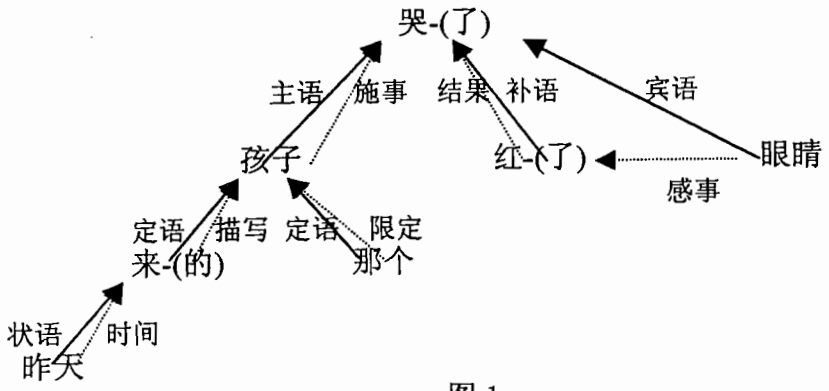

图 1

这种表达方式是一种依存关系的思路，其中带有方向的实线表示成分之间的句法关系（为了便于叙述，我们在这里借用常用的语法成分名称），它是语言成分在句子结构中的位置及其性质；带有方向的虚线表示成分之间的语义关系（也叫作事件角色），它是语言成分在句子中所对应的概念之间的联系。我们的实验系统把成分之间的句法关系和语义关系合在一起，作为句子的内部结构。这样的句子结构基本上能够表达图 1 的汉语句子与其英语译文之间必要的对比差异，可以避免输出不合格的英语译文，例如："somebody cried red (one's) eyes."或者"somebody cried (one's) eyes red."。

图 1 又不完全是传统意义上的依存关系树。它在以下几个方面有所扩充：

A）常见的依存结构表示的大多是句法形式和句法关系。依存语法认为，从属词的语义附加在它所依附的支配词的语义上，句子的语义关系可以由依存结构推断或变换而来（刘海涛，1997）。但是汉语句子有许多句法关系与语义关系不同构的现象，比如在上面那个句子中，"眼睛"在句法上受动词的支配，在语义上却与补语有感事关系（Experiencer）。为了表示这种现象，我们的依存关系树用两种联接结点的有向弧，分别表示句法和语义关系。对于同一个结点来说，两种弧的走向有时候会各不相同。

B）依存语法认为动词是整个句子结构的核心。我国学者曾经把依存语法定义为"主要研究以谓词为中心构句时由深层语义结构映现为表层句法结构的状况及条件，及谓词与体词之间的同现关系的一种结构语法"（周国光，见沈阳等，1995）。我们的实验系统也按这种观点在分析句子时重点考察动词对体词的支配关系。与此同时，我们还注意分析汉语句子中其他的依存关系，比如与形容词、名词等有关的句法和语义关系。

C）与传统的依存语法模型相同，我们表示句子结构的依存关系树不存在短语结点，也是多叉、多标记的。与其不同的是，我们的结点标记还可以反映句子中的语序。

经过这样扩充的树结构表示了句子的一种广义依存关系，它包括以下内容：

1）语言成分在句子中通过依存关系相互关联，依存关系包括句法依存和语义依存。

2）依存关系具有方向性。根据其方向，可以划分出成分之间的层级关系：上位成分、下位成分和同位成分。

3）一个上位成分可以带若干个下位成分；而一个下位成分只能从属于一个上位成分。

4）句法依存分为支配关系和附加关系两种，统称为直接联系。支配成分对被支配成分的选择限制来源于其词汇意义，后者是前者所表示的事物在概念上关涉的对象，它们共同表达该事物的完整的概念。附加成分与其上位成分的关系是，一个概念的范围由另一个概念加以限定。

5）语义依存是成分所对应的概念之间在意义上的联系。一个句子中有句法依存关系的成分之间不一定有语义依存关系，反之亦然。

我们的实验系统试图用这种广义依存关系来表示汉语句子的内部结构，系统分析的结果是输出句子的广义依存关系复杂特征集和树形图。

2 系统概况②

实验系统是基于规则的，系统包括分析控制程序、语言信息词典和语法规则库等部分。其中分析控制程序部分是整个系统的核心。它有两个功能，一是控制自动分析的操作过程，二是对规则进行识别、匹配和运算。语言信息词典和语法规则库为控制部分提供分析句子时需要的各种信息，前者存放与字词有关的信息，叫作词汇规则（Lexical rules），后者存放与句子结构有关的信息，叫作语法规则（Grammatical rules）。一个句子输入系统后，根据词汇驱动、词汇规则与语法规则相结合的分析策略，先从字构词，再造句，在构词造句的过程中识别出句子中的各个结构单位，分析出它们之间的句法关系和语义关系，用特征集合矩阵或者树形图表示出来。

这个实验系统的语言信息词典目前有大约 1300 个记录，语法规则库有大约 400 条规则。可以分析现代汉语陈述句一些常见的单句共约 600 句，主要是动词谓语句，包括及物动词的单宾语句、双宾语句、动补结构句、动词宾语句、形容词宾语句、小句宾语句、兼语式和连动式等句式。

3 构词

分析汉语句子首先遇到的问题是如何确定句子的基本结构单位。人们通常把词作为句法分析的起始点，按先构词、再句法的顺序分析汉语句子。构词法研究的是词的内部构造，以语素作为基本单位；句法研究的是句子的内部构造，以词作为基本单位。可是在汉语中，语素和词、词和词组、词组和句子，相互之间没有清楚的界限。在分析句子之前先分词，要解决的是构词问题，但却经常不可避免地遇到句法问题。有的学者认为这是一种语法层次的模糊性（刘群等，1998）。构词法和句法相互交错也是一个先理解还是先分词的问题。先有词，再分析，在处理以印欧语言为代表的许多语言时，几乎是天经地义的。可是在计算机里，汉语句子中的词不是现成的，需要一个一个地切出来，切分的过程和结果有时与理解句子无关，有时又常常与理解有关。当与理解有关时，我们就遇到了一个两难的问题：分词要以分析为前提，分析又要以分词为前提。按先构词、再句法的顺序分析汉语句子，会使我们很难避开这个难题。

如果考察人理解汉语句子的过程，我们不难看到，识别词和理解句子并不是截然分开的。以汉语为母语的人凭借语感能够识别绝大部分词，汉语自动分析虽然无法借助人的语感，但是可以借鉴人在识别词和理解句子时那种一体化的方式。也就是说，让计算机把字作为分析句子的起始点，用同一套算法，在同一个过程里，一面从字串中识别分析句子所需要的基本结构单位，一面以这些结构单位为基础分析句子的句法和语义关系。（在汉英机器翻译研究中已有学者针对汉/英两种语言词语层次不平行的问题，提出这种设想（刘群等，1998））。这些基本结构单位相当于通常所说的词或词组。在汉语自动分析系统中，它们是句法和语义分析的基本单位。如果用在机器翻译中，它们还应该体现源语转换成译语时必要的对比差异。

根据这种思路，我们在汉语分析系统中没有单独设立自动分词阶段，由字组词和由词构句都集成在同一个分析过程中。构词时需要区分词或词组的不同类型（见图 2）。

其中第 1 类是单纯词和不可扩展的复合词（包括大部分偏正、主谓、联合、重叠结构、附加结构，和少量动宾、动补结构）。这些词需要直接收入词典。

第 2 类主要是离合词，可分为动宾和动补两种结构。其中动宾结构复合词又分为 5、6 两类，动补结构复合词也分为 7、8（动结式和动趋式）两种，

第 5 类是限制性扩展（如“洗过澡”、“洗没洗澡”、“洗一回澡”）。可以归纳其所有的扩展格式，用动宾式构词规则处理。

第 6 类是自由扩展（如“洗了一个热水澡”、“出了一身虚汗”）。按单宾语句式用句法规则处理。

第 7 类是动结式，它的自由扩展形式是“得”字句（“吃得比昨天饱”），在句法分析

时处理；它的限制性扩展则按动补句式由句法规则处理。

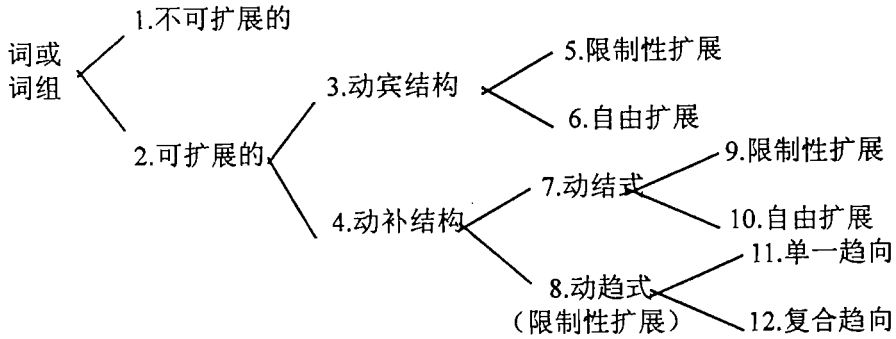

图 2

第 8 类动趋式只有限制性扩展，其中单一趋向补语（“提不出”、“提得出”类）由动趋式构词规则处理（包括意义虚化，即表示结果和状态的情形），复合趋向补语（“提出来”、“提起来”、“提了出来”）需要先合成复合趋向动词（“出来”、“出去”、“起来”等等），然后与单一趋向补语同样处理。对于可带宾语的动趋式，再转句法规则处理。

扩展的复合词经过构词规则的处理，可以还原成基本词形，同时分离出动量成分、体成分、否定成分和各种语义关系信息（结果、状态、可能等），标记在该词的结点上。构词的处理方法可参见（傅爱平，1999）。

4 分析句子

系统分析句子的过程是双向扫描，由词汇规则驱动。其中左向扫描是从句末向句首，右向扫描正好相反。分析从左向扫描开始，逐字进行构词和构句分析，到句首或到规定的折返点后再返回作右向扫描分析。

左向扫描主要处理成分之间的支配关系。据考察，汉语句子中词语的支配关系，尤其是动词的支配关系，其典型的语序表现大多是支配成分在前，被支配成分在后；附加关系则多表现为附加成分在前，中心成分在后（徐通锵，2001 提到了同样的观点）。从句末向句首分析，可以先处理位于后面的被支配成分。为分析支配关系准备条件。右向扫描要处理的主要是附加关系和非典型语序的支配关系。右向扫描如果是在规定的折返点开始的，要在分析结束时再回到折返点，继续原来的过程。依存关系首先是一种结构层次关系。这样的处理流程能够体现层次分析的思想，在自底向上归约句子时，得到用下位成分、上位成分、同位成分等表示的结构层次关系。

系统用两种规则分析句子：词汇规则和语法规则。前者描写词语的使用情况，以构词和支配关系为主，后者描写构句的规律，附加关系规则属于这一类。分析句子的时候，系统先调用词汇规则，后调用语法规则，以词汇规则为主，以语法规则为辅，这就是词汇规则驱动的意思。

我们在英汉机器翻译系统中也用词汇规则驱动作源语（英语）分析，主要目的是使分析的立足点具体化，以便提高分析的精确程度。对于汉语自动分析来说，词汇规则驱动还有一个作用，这就是在识别句子的依存结构时更多地依据词语的支配关系信息。人们常说，汉语按“意合”的方式生成句子。所谓“意合”，应该是指成分之间在意义上的联系，也就是我们所说的语义关系。怎样识别这些语义关系呢？象分析英语那样从结构形式入手，主要依据句法关系是远远不够的。这是因为，汉语句子的语法结构和语义角色之间的联系比较松散，句法成分和语义成分的配位很灵活，一个句法结构经常表示多种意义，一种语义关系也经常可以用多个句法结构来表示。另一方面，汉语句子的结构形式又有通过缩减趋

于简化的趋势，这就使得句子的结构关系和语义关系常常不同构。所以分析汉语句子最好尽可能多地依靠词语在概念上的相互关联，支配关系信息就是其中之一。

我们以动词的规则为例说明词汇规则的作用。动词以义项为单位建立支配关系规则，描写被支配成分的类别（句法范畴、语义特征等）、整个支配关系格局的典型语序表现和句法、语义关系。例如动词“罚”，支配成分及其类别分别是：动作的主体 N0（一般是人或机构），动作所涉及的直接客体 N1（一般是人或机构），以及动作涉及的间接客体 N2（一般是财物）。在句子表层，典型语序和句法、语义关系是：

N0（主语/施事） + “罚” + N1（宾语 1/与事） + N2（宾语 2/受事）

动词的支配关系规则把支配成分的类别和典型语序作为识别条件，句法和语义关系是规则的结论。根据条件可以识别句子中与典型语序一致的情况，对于不一致的情况，需要调用变换分析规则处理。

变换分析规则是语法规则当中的一类，用来处理相同的支配结构在句子表层的各种变换。如上面的规则 N0+V+N1+N2，可以有下面几种说法：“警察罚了司机钱”，“警察从司机那儿罚了钱”，“司机被警察罚了钱”，“司机被罚了钱”，还有各种转指的“的”字结构：“被警察罚了钱的司机”，“罚司机钱的警察”，“被罚了钱的司机”，“司机被罚的那笔钱”。虽然在这一组句子中，N0、N1 和 N2 出现在不同的位置上，有时甚至不出现，但是它们与动词的语义关系是不变的。我们的系统把这些句式看作基式 N0+V+N1+N2 的各种变换形式。基式在语言信息词典里用支配关系规则表示，变换形式在语法规则库里用一组变换分析规则描写。支配关系规则是一种个性规则，它因词语及词语依存结构内的语义关系不同而异。变换分析规则是一种共性规则，它描写的是同一组语义关系由于句法的规定性而在句子表层表现出来的不同句式。在使用基式作变换分析的时候，语义特征分析是不可缺少的手段，因为识别支配关系常常要借助被支配成分的语义类别，细致地考察每个成分的语法范畴和语义特征。变换分析规则也常常要借助语义特征区别不同的变换关系，例如同是 N0+V+N1+N2 中的动词，“给予”类和“获取”类的变换关系就不完全相同。

除了描写支配关系，词汇规则也描写构词（包括各种扩展方式构词）、成语、固定搭配、以及局部限定关系等信息，虚词的词汇规则还描写构句信息。

由于篇幅所限，略去说明句子分析过程的例子。

5 与句法关系不同构的语义关系

依存结构中的句法关系和语义关系大多是同构的。也就是说，相互依存的成分具有相同的句法关系和语义关系指向。但在分析汉语句子时也常遇到句法和语义不同构的现象，比如下面两种情形。

第一种是句法上不具有直接联系的成分，相互之间存在语义关系。比如图 1，在该句的动补结构“哭红了眼睛”中，“眼睛”与“红”就是这种关系。我们的系统除了能识别出动词与其补语和宾语的依存关系以外，也能识别“眼睛”和“红”之间的语义关系：前者是后者的感事。对于句子“学生们看懂了这本书”，系统除了输出“看”与“学生们”（主语/施事）、“懂”（补语/结果）、“这本书”（宾语/受事）的依存关系以外，还可以给出“懂”与“学生们”（施事）和“这本书”（内容）的之间语义关系，它们也是在句法上没有直接联系的成分。

另一种情况是成分之间句法联系和语义联系的方向不一样。比如在上一节的例句中，动词“罚”与其支配成分组成的“VP 的”短语在句法上是“警察”的修饰成分。而在语义上，“VP 的”短语转指“罚”的施事“警察”，“警察”又是“罚”的支配成分。所以在句法上“罚” 位于“警察”的下位，在语义上却正好相反。动词通过转指形式修饰它的被支配成分，这在汉语里很普遍。我们的系统在分析这种情况时，除了能得到动词短语内部的句法关系和语义关系、动词短语对“的”字后面的名词的修饰关系以外，还可以得出动词短语与被它修饰的名词之间的语义关系（“施事”、“受事”、“与事”等等）。在汉英机器翻

译的转换生成中，这些信息有助于选择英语描写性修饰成分的句法表现和形态表现。

这些与句法关系不同构的语义关系在汉语语法研究中常被叫作语义指向，它们表示句子深层的逻辑或意义关系，能够帮助计算机更好地理解句子的意思。如果用在汉英机器翻译中，会有助于表达汉语句子与其英语译文之间必要的对比差异，有助于生成比较贴切的英语译文。

6 有待解决的问题

系统在实验过程中也遇到不少问题。例如，支配关系应该是广义的，即除了动词以外，一部分名词和形容词也有支配关系。利用名词和形容词的支配关系分析句子，会更有助于顺应汉语“意合”的特点。问题是如何在系统中形式化地表达这一类支配关系，它们在分析算法上与动词有什么不同。目前我们的系统还没有解决这个问题。即使是动词的支配关系，目前的处理方法也还不能令人满意。比如动词“发表”，在“登载”这个义项下，被支配成分是作者 N0（一般是人或机构）、作品 N1（一般是文字作品）和载体 N2（一般是出版物）。在典型情况下，N0 的句法位置是主语，N1 是宾语，N2 是动词的修饰成分（状语）并被介宾化。例如“老李在人民日报上发表了一篇文章”。但是“人民日报发表了老李的一篇文章”和“人民日报发表了那篇文章”也很常见。这时 N2 出现在句子表层的主语位置上，N0 和 N1 还可以表现为一种形式上的领属关系。这些变换有什么规律？变换的条件是什么？如何把它们形式化地表达出来并且集成到系统中以便操作？目前我们对这一类问题还没有比较系统的研究。

另外，系统采用的词汇主义的分析策略（词汇驱动、词汇规则和语法规则相结合）相对注重句子的局部分析，如果没有把两种规则恰当地衔接起来，可能会造成局部分析错误或者全句的结构脱节。我们在分析英语时曾经遇到个别这样的例子。汉语句子受结构形式的制约比英语要少得多，因此出现这类问题的可能性还会大。现在系统分析的都是单句，数量也不多，基本上没有发现这类错误，但还不足为凭。

此外，目前定义的广义依存关系能够在多大程度上表达汉语的各种句式，覆盖多少汉语的语言现象，目前的分析算法能否处理得了更多的语料，在系统还没有经过相当多句子尤其是相当多复句的检验时，也还不能轻易断定。

注：
①系统调试环境的改造得到了刘倬先生的具体指导，谨此致谢。
②研制工作是在中国社会科学院基础研究课题资助下进行的。张翊、胡凤国参加了部分程序设计的工作。

参考文献

[1] 冯志伟 现代语言学流派，陕西人民出版社，1987
[2] 傅爱平 汉英机器翻译源语分析中词的识别，中文信息学报，1999（5）
[3] 胡明扬 （主编）西方语言学名著选读，中国人民大学出版社，1988
[4] 刘海涛 依存语法和机器翻译，语言文字应用，1997（3）
[5] 刘 群等 汉英机器翻译的难点分析，1998 中文信息处理国际会议论文集，清华大学出版社，1998
[6] 刘 倬等 JFY-IV 机器翻译系统概要，中文信息学报，1989（4）
[7] 陆志韦 汉语的构词法，科学出版社，1957
[8] 沈 阳等（主编） 现代汉语配价语法研究，北京大学出版社，1995
[9] 王海峰等 汉英机器翻译中汉语离合词的处理策略，情报学报，1999（4）
[10] 徐通锵 字和汉语语义句法的基本结构原理，语言文字应用，2001（1）