一个实用化的俄汉机器翻译系统

侯 敏　孙建军　陈英琦　薛选民　侯 方　毕德刚

（黑龙江大学　哈尔滨　150080）

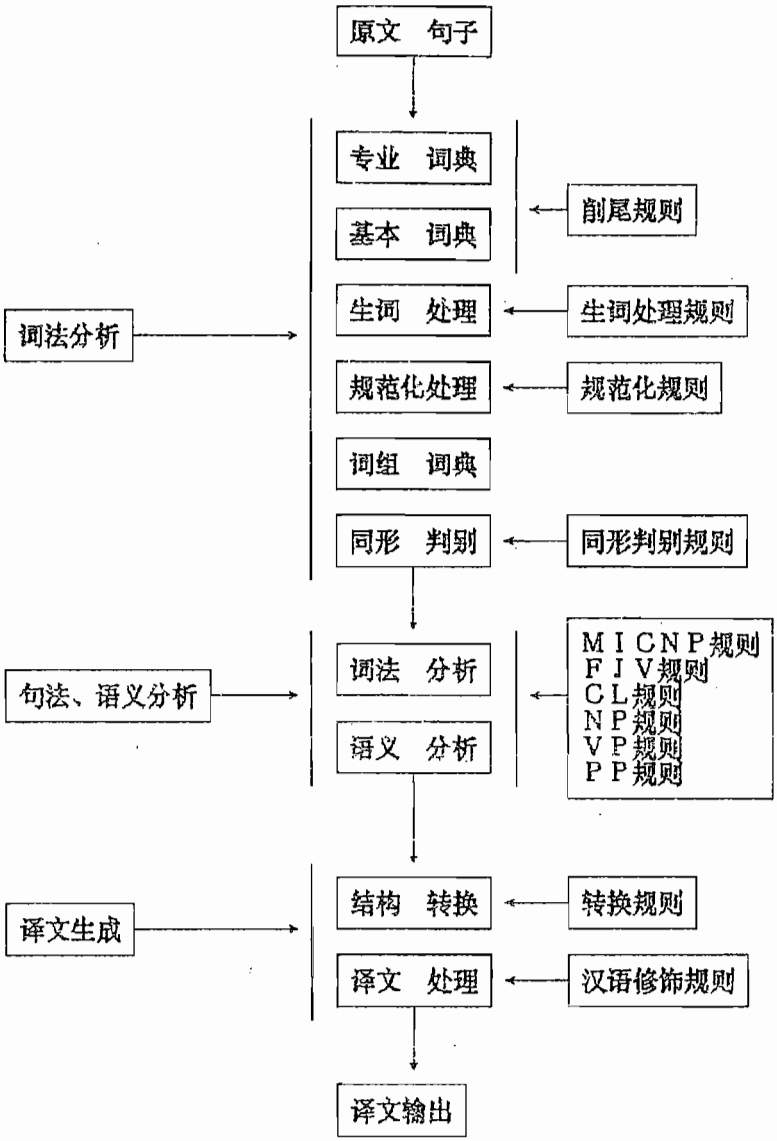
摘要：全译通俄汉机器翻译系统是一个实用化的全自动的电脑翻译系统。系统包括基本词典六万词条、专业词典(经贸方面的)二万词条、词组词典八千余条，语言翻译规则一千八百余条。全部程序用 C 语言编写，语言规则用专门设计的规则描述语言RDL书写。系统以句子模式匹配技术为核心，句法上采用自顶向下和自底向上相结合的分析方法，并且在不同平面上分别解决多义识别问题。

A PRACTICAL RUSSIAN-CHINESE MT SYSTEM

Hou Min　Sun Jianjun　Chen Yingqi　Xue Xuanmin
Hou Fang　Bi Degang
(Heilongjiang University　Harbin　150080)

The TRANSYSTEM Russian-Chinese MT system is an operational and fully automatic one. The system consists of a basic dictionary(60,000 entries), a technical term dictionary (concerning economy and trade, 20,000 entries), a phrase dictionary(8,000 entries), and more than 1,800 linguistic translation rules. All the programs are written in C, and the linguistic rules in RDL, which is a specially-designed Rule Description Language. The system takes the technique of pattern matching as its core. In syntactic analysis, the method of combining top-down and bottom-up parsing is adopted, and problems of polysemy are dealt with respectively at different syntactic levels.

"全译通俄汉机器翻译系统"是一个应用型的全自动的电脑翻译系统。该系统日前已通过省科委组织的专家鉴定。系统包括电子词典(基本词典六万词条、专业词典(经贸方面的)二万词条、词组词典八千余条)、语言翻译规则(共一千八百余条)、以及为实现翻译加工过程而编写的程序。整个软件系统的流程，从原文输入到译文输出，要经历查词典、生词处理、规范化处理、同形判别、句法分析、语义分析、结构转换、译文处理等步骤。从语言分析的角度看，分析是以句子为单位来进行的。通过对原文进行词法、句法、语义等多层次的分析，得到一个原文句子的多结点的带有语义标记的句法树结构。然后再根据源语言（俄语）和目标语（汉语）的对比分析，并按照目标语的语法规律，把原文的树结构转换成相应的译文的线性结构，从而生成译文句子。显然，整个分析过程，也是自始至终不断应用各类规则的过程。它们之间的相互关系，可用下面框图表示：



下面就系统的一些问题分别作简要介绍。

一、电子词典

我们建立了基本词典、词组词典和专业词典等几种电子词典。目前，专业词典还只有经贸专业领域一种。

俄语是综合型语言。根据俄语的特点，为节省存贮空间和提高检索效率，在建立词典时，对俄文中没有形态变化的词，如副词、连词、前置词等，我们存入该词本身；但对有形态变化的词来说，如动词、名词、形容词，我们只把其原形或词干存放在词典中。系统查词典时，应先按照“削尾规则”将该单词恢复为其原形或词干，然后再在电子词典中查找。不规则变化的词则应将各种特殊形态变化的词形都存放在词典中。

为加快查找速度，我们采用建立索引文件帮助查找的方法。索引文件的建立是采取分类与B树技术相结合的方法实现的。实践证明，建立的索引文件是高效的。用C语言编程，其查找速度在286机上为每秒35个词，（词典容量为6万词条左右）。与此同时，由于采用了B树技术，整个词典在工作过程中仅占几K内存空间，克服了一般词典空间开销大的缺点。

为方便用户参与开发扩展本系统，我们还专门设计了用户词典。用户只需具有初步俄语知识，就能独立胜任工作，根据需要，利用这部词典，随时增添词条。其具体做法是，由用户给出所要增添词的索引词、词性以及汉义（如有可能，给出与该词用法相近的参照词更好），系统即能自动填补其他必要的信息。用户词典主要用来增加系统词典中缺少的词，也可用来改变系统词典内中已有词的汉义。例如，用户在一篇经系统翻译加工的文章中，对某个词的汉语译法不满意，则可在用户词典中再存放该词，并给出所要求的汉义，从而得到满意的译文。如果这一修改只是一时性的需要，事后可将该词条在用户词典中删去。可以看出，用户词典在诸多词典中是最为优先的。

二、词法分析

§1．查词典

查词典的目的在于获取词的信息，无形态变化的词的信息全部存放在词典中，而有形态变化词的部分信息，如名词、形容词的数、格，动词的时态、人称等属性则需通过应用削尾规则才能随机给出。

§2．生词处理

生词指的是没有存入在词典里的词，或者说，是经过查词典查不到的词。生词处理这一环节之所以必要，不仅在于词典可能不完善，会有遗漏；也不仅在于自然语言在演变过程中会不断产生新词，这些只是问题的一个方面。而更重要的，还在于在具体翻译实践中，往往会遇到一些不成其为词的词，如象有些商标名、不常用的缩写词、人名、地名、符号等。这些词，在人用的词典里没有把它收录进去，当然也不可能把它们全部编到机器词典里去。如果可以把前面提到的漏词、新词看成是临时性生词的话，那么，后面的那些就是固定性生词。它们将永远存在。对生词的处理需要根据词的形态来判定其词性及数、格等，以便参与分析加工。如词尾为ый、ая、ое等的，可断定为形容词、单数、第一格；词尾为ому、ему的，可断定为形容词、单数、第三格；词尾为ами、ями的，可断定为名词、复数、第五格。至于汉义，可以词原样给出。这样做虽说是不得已而为之，但对那些固定性生词来说，却不失为一种合适的求解。

§3．规范化处理

规范是就系统进行翻译加工是否方便而言的。规范化处理就是把系统难以加工的结构，施用某种手段，调整为系统易于分析的形式。俄语句子中，有时会出现某个成分被省略，或者在句子中间插入了一些词语，或者正常的词序颠倒了等情况。这些现象给分析带来很大困难。设法把省略的成分找出补上，插入的词语挪置一边，颠倒的顺序正过来，就能顺利匹配规则。俄语中有时还会出现结构分离现象，例如在有的从句中，前置词往往要随关联词一起移到从句句首。这时，如果该前置词同时又恰好应该与从句动词组合成动词短语，那么它们之间的密切联系在线性结构中就被分离开了。如“Мне не совсем ясно, над кем вы смеётесь.（我不很清楚，您在笑话谁。）”句中动词“смеяться”是“笑”的意思，当它和前置词组合在一起“смеяться над（кем～чем）”才是“嘲笑”“笑话”的意思。然而现在动词不仅与前置词颠倒了顺序，而且中间被隔断了。处理这类问题，我们反向应用乔姆斯基的“踪迹理论”（Trace Theory），把前置词及其支配的关联词一起移到从句动词后，也就是从表层结构形式又返回到深层结构。移走的地方留下踪迹，作为从句的标记。这样，动词和前置词的密切联系得到恢复，而作为从句标志的关联词的踪迹仍然存在，分析问题也就迎刃而解了。

§4．同形判别

俄语中有些词类是可以互相演变转化的。如，形容词可以演化为名词，больной，是形容词“有病的”，但又用作名词“病人”；副词可以演化为前置词，впереди，作副词表示“在前面”，作前置词表示“在…前面”。这种词，我们把它叫做兼类词或跨类词。跨类词可以有两个或更多的词性，但是，它们在一定的语言环境中只能有一个确定的词性，如“впереди”在“Он шёл впереди.”（他走在前面）中是副词。在“Он шёл впереди всех.”（他走在大家的前面）中是前置词。这种跨类词是语言本身具有的。另外，由于有形态变化的词我们在词典中是以词干或原形的方式存入的，所以有些原本不是跨类的词也变得同形了。如добро是名词，добрый是形容词，词形本不同，但削去尾以后，都是добр，在词典中也成了“跨类词”。还有，俄语形态变化丰富，有些不相干的词，变化后却面貌相同了。如“ряд”是

名词，可它的单数第五格形式却与副词“рядом”同形，于是也成了“跨类词”。要想顺利翻译句子，就需要对这些词作同形判别。同形判别就是根据词在句子中的前后环境条件，来确定它的词性归属，以便正确给出词性及汉义，参与句子分析。例如，判断当前词是副词还是前置词要根据下面的同形判别规则：

```
*/V|C|,|EW=:F
*=:P
```

这条规则中，* 号表示当前词，如果它后面是动词、或连词、或逗号、或句末标记，那么这个词就判定为副词，否则就判定为前置词。

三、句法分析

§ 1．规则描述语言

规则描述语言是我机器翻译研究所在十几年机器翻译实践中研究并建立的一套形式化的描述语言。建立规则描述语言是为了保证语言规则和程序设计完全分开，从而真正做到语言规则可以任意增、删、改而无须更动程序。

规则描述语言要根据各个机译系统进行语言分析综合所需要的种种功能来设计。我们主要采用模式匹配技术，要求具有绝对匹配、或匹配、可有可无匹配、预示匹配等逻辑匹配功能，以及调序、增词，删词、归结、分枝等操作功能。

§ 2．句法规则库

句法规则库主要由以下几个子规则库组成：

1）微型名词短语规则（ MICNP ）

这类规则把名词前可能出现的修饰语，如数词、形容词等，与该名词合并在一起，组成一个初步的微型名词短语。

2）并列结构规则（ FJV ）

并列结构规则主要处理并列的副词、并列的形容词，或并列的动词等，不涉及并列句的处理。

3）句子结构规则（ CL ）

句子结构指的是句子最顶层的框架结构。CL 规则除了处理一般句型外，还包括一些特殊的句型。

4）名词短语规则（ NP ）

NP 规则是在 MICNP 规则的基础上，把微型名词短语再进一步加以扩展，也就是把名词后可能出现的该名词的修饰语，如二格名词、前置词短语、不定式短语、定语从句等，把它们再和微型名词短语合并，组成一个完整的名词短语。

5）动词短语规则（ VP ）

动词短语规则是说明动词用法的规则。我们根据动词的不同类别如系动词、不及物动词、及物动词等，将动词分为V1 、 V2 、 V3 、 V4 、 V5 等类，其中有些类的动词如及物动词等，还可以再细分为若干小类，如 V3A 、 V3B 、 V3C 、 V3D、……等。

6）前置词短语规则（ PP ）

俄语前置词短语中，前置词的含义往往随着它所支配的名词短语的不同而不同。因此，前置词短语规则的制订必须落实到每一个具体的前置词上。它们实质上是一些特殊词的词处理规则，每一个前置词都有它自己的若干条规则。PP 规则只是这些子规则库的总称。

§ 3．句法分析过程

分析是以句子为单位进行的。

句法分析的过程是不断应用句法规则的过程，也是逐步进行模式匹配的过程。

应用句法规则时，在同一规则库中，不同的规则之间是“或”的关系，即只要有一条规则执行成功，就认为该规则库的执行是成功的。在同一规则中，扫描动作中各个匹配动作之间是“与”的关系。它们依次匹配待处理结点序列中的结点，当所有匹配动作都匹配成功后才执行生成动作结点，并认为该规则的执行是成功的。同时，这也意味着某一句法结构模式匹配成功。

一个句子的句法分析，一般说来，要经历两次扫描，先是一次自句尾向句首进行的扫描，使用的规则是 NICNP 规则和 FJV 规则，目的在于简化结构，以便下一步使用 CL 规则。之后是一次自句首向句尾进行的扫描。主要应用 CL 规则，其余的规则，如 NP 规则、VP 规则、PP 规则等，是在执行 CL规则过程中，根据需要随机调用的。这两次扫描有一个共同的目的，那就是：建立这个句子的一棵句法树。它们之间的不同在于：从扫描方向来看，第一次扫描是自句尾向句首方向进行的扫描，而第二次则是自句首向句尾的。从主次关系来看，第一次扫描是为第二次扫描打基础服务的。从采用的方法来看，第一次扫描采用的是面向数据的自底向上的分析方法，而第二次则是基于假设的自顶向下的分析方法。

四. 语义分析

§1. 多义问题的处理

要想提高翻译质量，仅仅做句法结构的分析是完全不够的，还必须深入到句子的内部进行语义分析。在语义问题上，对机器翻译来说，最难的莫过于多义识别了。根据多义的不同性质，我们从不同角度在不同平面上进行了处理。

1）词汇多义

有些词汇多义问题，在进行词法分析时已经解决。如某些跨类词，当某一类词用时一个意思，当另一类词用时是另一个意思。但经过同形判别，词性判定，词义也随之确定。另外，有的动词，既可作及物动词用，也可作不及物动词用，而且汉义也因此不同。这时可根据它们用法不同，匹配不同的句法规则，从而得出不同的汉义。以上是语义问题转化为词法或句法问题处理的情况。

一个词，当某类词或某个次类用时，也可能有几个义项。处理这类问题往往涉及词与词之间的搭配关系。例如，形容词的多义取决于它所修饰的名词。动词的多义则要看它所支配的补语。为了方便进行语义分析，我们把名词做了语义分类，给出每个名词的必要的语义信息。同时我们还要在多义词的词条中，给出每个义项的判定条件。这样就可以根据特定的上下文去选择并判定当前词的义项。

有时，一个多义词的词义要依据它自身条件来判定。如有的名词复数时一个意思，单数时又是一个意思。有的连词位置在句首和不在句首词义不一样。

多义词在特定组合中的意义则以词组的形式给出。

实在无法解决的多义词则把各个义项都给出。

2）结构多义

结构多义指的是词或短语和别的词或短语形成两种或两种以上的结构关系。例如一个同时具有二格和四格的名词，在结构上就既可能是它前面名词的定语，又可能和前面名词并列作某个动词的补语。结构多义问题多数在句法分析过程中可以得到解决。有的在局部无法解决，要进入更大的结构中才能处理。有的光用句法分析不行，还要结合语义分析。个别特殊情况要采用特殊对策。例如，“студенты и студентки этого университета”这是两个并列的名词后继一个二格名词，这个二格名词是只修饰后一个名词、还是同时修饰这两个名词呢？也就是说，这个短语的意思是“男大学生和这所大学的女大学生”还是“这所大学的男大学生和这所大学的女大学生”，这是个很难判定的问题。我们把它译成“这所大学的女大学生和男大学生”，问题就得到解决。因为这个汉语译文在结构上也是歧义的。在这里我们采用了以歧义对付歧义的策略。

3）前置词多义

前置词多义兼有词汇多义和结构多义二重性，一个前置词，支配不同的名词，含有不同的含义。例如，на＋时间名词，表示“在某时间”，на＋四格处所名词，有“到某处去”的意思，на＋六格处所名词，则表示“在某处”，на＋会议类名词，表示“参加某会议”的意思，“на＋食物名词”表示“吃某物”的意思……。由于前置词具有这种词汇多义的特性，我们对前置词的处理就

落实到每个具体的前置词上，把前置词作特殊词处理，每一个前置词都有它自己的若干条分析规则。另一方面，前置词及其所支配名词组成的前置词短语在句子中既可以作定语，也可以作状语。这类结构多义本是语言分析中的一个难点。人来分析，往往要应用客观世界知识，即使如此，有的问题依然束手无策。我们根据表达所属关系的多半用作定语，表达时间、地点等内容的多半用作状语等情况，把前置词短语进行分类，使问题得到部分解决。

§2．语义和句法的关系

系统是以句法为主设计的，语义问题随着句法分析的进展逐步进行处理。我们采取了在不同句法平面上分别处理各类语义问题的方法。例如，在句子结构平面上确定谓语动词如何根据不同主语来选取词义，在名词短语平面上选取形容词对中心名词的恰当汉义，在动词短语平面上解决动补的搭配等。

五、译文生成

§1．结构转换

我们在对源语言（俄语）和目标语（汉语）进行充分对比分析研究的基础上，按照目标语语法规律，采取了原文分析与译文综合同步完成的方法。也就是说，在建立原文树结构的同时，也生成了译文的链结构。我们句法规则的组成形式清楚地体现了这个分析与综合同步完成的过程。规则的模式部分是一个多项式，各个项表明原文树结构分枝的多个结点，树结构的结点名就是该规则所在规则库的名称。操作部分则规定了相应的译文的一个词序列。一旦某条规则匹配成功，则一棵原文的树和一条译文的链就同时产生。而且，由于我们语言分析的过程是递归调用各类规则的过程，这种树链结构转换也就始终贯穿在整个翻译加工的过程中，直到一个句子加工结束。我们的目的是翻译，分析原文只是达到目的的一种手段。原文的树结构随着建立就被扬弃了，留下的只是相应的汉语译文。

§2．汉语修饰

汉语没有形态变化，因此，一个单独的译文生成过程就显得没有必要。有关汉语生成需要做的一些主要工作，如调序、添词、删词等，在结构转换时已经完成。这里提出的汉语修饰是对生成的译文中可能出现的问题作进一步处理，如删除多余的“的”、“到”、“在”等，使其更符合汉语习惯。

六、系统的特点

§1．本系统是一个应用型的、全自动的俄汉机器翻译系统。系统除专业词典外，还提供了用户词典，以便用户直接参与开发。此外，系统采取生词处理、规范化处理等措施，完全排除译前、译后加工等人为干预因素，做到彻底的全自动。我们认为，机器翻译本质上是对人的翻译能力的模拟。机译与人译相比，机译的质量不可能赶上人译。机译的长处在于速度快，费用低。要想保持这个优势，全自动是完全必要的。

§2．系统采用通用的机器翻译专用软件，加快了研制开发的速度。这套机译专用软件，是在建立规则描述语言、统一规定词典格式等基础上，并经过英汉、日汉两套机译系统试用，逐步完善起来的。现在又经历了俄汉机译系统的经验。事实证明，用它来搞外汉机器翻译是十分方便的，能大大缩短研制开发的周期。

§3．系统在句法分析方面采用自底向上和自顶向下相结合的分析方法。前者是以数据为基础的，通过归结函数实现语法树的建立。后者是从假设出发的，通过预示匹配函数，递归调用各类规则，达到逐层分支的目的。我们先进行自底向上分析，扫清树叶，露出较平后再进行自顶向下的预示分析。这种分析方法回溯少，匹配成功率高，加快了翻译速度。

§4．系统具有在各个不同平面上根据语境条件选取汉义的功能。例如，在S平面上确定主谓搭配汉义的选取，NP平面上选取修饰语对中心名词的恰当词义，在VP平面上解决动补的搭配关系等。前置词的句法功能和搭配多义则通过语义分类和具体词规则等来解决。

§5．本系统是一个以模式匹配技术为核心的系统。每条规则都含有一个要匹配的模式，即判断语句。模式不以其匹配对象的不同情况进行分类，死板地规定几种固定格式，而是只有一个统一的格式，即所有模式都是一至任意多个逻辑语句的自由组合。这样做一方面使用者可以随心所欲地来组合各种模式，另一方面由于模式的组合形式是无限的，就能充分反映自然语言的复杂多变情况，既能自如地对自然语言进行结构描写，又便于把多种结构抽象概括在一个模式里，以达到以简驭繁的目的。

参 考 文 献

[1] Chomcky ,N. , Lectures on Government and Binding. Fris Publications, Dordrecht.1981.

[2] Jakendoff,R. , Semantics and Cognition. MIT Press. 1983.

[3] Kondo,T. , Enumeration of Sentence Types of Language and Its Relevance to Machine Translation. In : Language and Artificial Intelligence, ed. by M. Nagao, PP.303-323. 1987.

[4] Papegaaij,B.C. , Word Expert Semantics — an Interlingual Knowledged-Based Approach Foris Publications,Dordrecht. 1986.

[5] 黑龙江大学机器翻译研究所， TRANSYSTEM 俄汉机器翻译系统技术报告. 1993.