机器翻译与语言研究*

詹卫东+ 常宝宝++ 俞士汶++
{zwd,chbb,yusw}@pku.edu.cn http://icl.pku.edu.cn/
+北京大学中文系 ++北京大学计算语言学研究所

提 要 本文从机器翻译的一般模式谈起，讨论了服务于机器翻译的语言研究应该注意的问题，包括：应更加重视形式与意义之间对应关系的系统研究；应重视以机器为“标准”来扩大关注的语言现象的范围；应重视在形式化的知识表示框架下进行大规模的语言工程实践的研究工作。

关键词 机器翻译 语言研究 电子词典 短语结构规则 形式化 语言知识库

一 什么是机器翻译？

机器翻译（Machine Translation）指的是利用计算机程序把一种语言的文本（可以称为源语言文本）翻译成另外一种语言的文本（目标语言文本）。这项研究的意义是不言而喻的。很久以来，人们就梦想着有朝一日，能造出一种设备清除人类交流过程中的“语言障碍”，使得使用不同语言的人能自由地相互交流[1]。在当代信息社会，语言障碍的问题更加突出。大量的政府文件、商业以及科技资料都需要在短时期内得到翻译，互联网的问世更是扩大了翻译需求。可以说，人们现在比以往任何时候都迫切希望拥有自动翻译技术。然而，过去五十多年机器翻译的研究历史却表明，机器翻译的困难程度和复杂程度远远超出了最初倡导机器翻译研究的先驱者们的想象。机器翻译至今仍是一项十分具有挑战性的研究课题。其进展不仅需要计算手段的创新，更要依赖于人们对语言本质以及语言计算模型认识的进展。可以说，语言学研究的水平对机器翻译系统研制的成败起着十分关键的作用。出于这样的认识，本文将讨论机器翻译对语言研究的要求，希望吸引更多语言学研究者将机器翻译作为思考语言学问题的一个参照系，使更多的语言研究成果，可以为机器翻译提供帮助。

机器翻译系统的研制工作从上个世纪四十年代末开始，至今已经发展出许多不同的方法。总体来看，现有的机器翻译方法可以归纳为三种类型，一种是基于规则（rule-based）的方法；第二种是基于统计（statistic-based）的方法；第三种是基于实例（example-based）的方法。限于篇幅，这里我们主要介绍第一种方法的工作模式。

基于规则的机器翻译方法把翻译过程看作是一个在语言学知识引导下的符号变换过程。这种方法要求把有关源语言和目标语言的知识以计算机可以操作(“看懂”)的形式表示出来。下面以汉英机器翻译为例说明翻译的基本过程：

（1）对源语言进行词法分析，这个阶段利用源语言词汇层面的知识，识别出源语言文本字符串中的单词，并从词典中获得每个单词的句法语义知识，以备在后续处理中使用。

例如：汉语句子“她把一束花放在桌上”，经过词法分析，会得到下面的结果[2]：

她/r 把/p 一/m 束/q 花/n 放/v 在/p 桌/ng 上/f 。/w

* 本文研究工作得到“高等学校全国优秀博士学位论文作者专项资金”和国家 973 课题“面向新闻领域的汉英机器翻译系统”（项目号：G1998030507-4）资助，特此致谢。

[1] 可以访问著名的《连线》（WIRED）杂志网站 http://www.wired.com/wired/archive/8.05/timeline.html。这篇文章把机器翻译的理想上溯到 1629 法国数学家兼哲学家笛卡儿（René Descartes）的时代。

[2] 斜杠后字母是词性标记，r 表示代词，p 介词，m 数词，q 量词，n 名词，ng 名词性语素，f 方位词，w 标点符号（下文所用标记符号含义同此）。

（2）对源语言词串进行句法分析，得到句法结构以及跟结构相关的特征结构（feature structure）。这个阶段的处理需要用到句法层面的知识，通常是表示为扩展的上下文无关规则（本文第二节有更多的说明），由两个部分组成，一个部分是上下文无关规则（Context Free Rule），指明了短语的组成关系，例如：一个名词短语可以由一个数量结构和另外一个名词短语组成这个事实用上下文无关规则可以描述为：*np→mp np*。另外一个部分是一组“合一等式”（unification formula），主要描写在什么条件下可以用这条规则进行组合以及组合之后得到的新的语言单位的属性信息（参见詹卫东 2000）。句法分析的结果可以表示为一棵句法树。例如对上述汉语句子进行句法分析，将会得到图-1 所示的句法树[3]（这里略去了特征结构信息）。

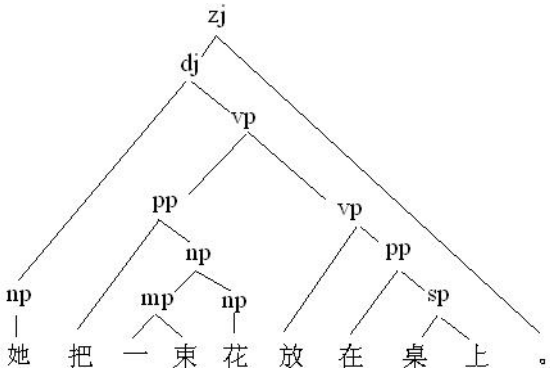


图-1

（3）把源语言的句法结构转换为目标语言的句法结构，结构转换主要利用源语言结构和目标语言结构之间的对应关系进行，通过一组转换规则来指导把源语言的句法树转换成目标语言的句法树。转换规则列出了源语言的句法结构以及对应的目标语言结构，并描述了这种转换关系成立的条件，对于上述例子，图-2 描述了这种结构转换前后的对应关系，图中右部是转换后得到英语结构树，其中每个树结点都标有两个以斜线分隔的范畴标记，斜线左边的范畴是由斜线右边的范畴转换得到的。

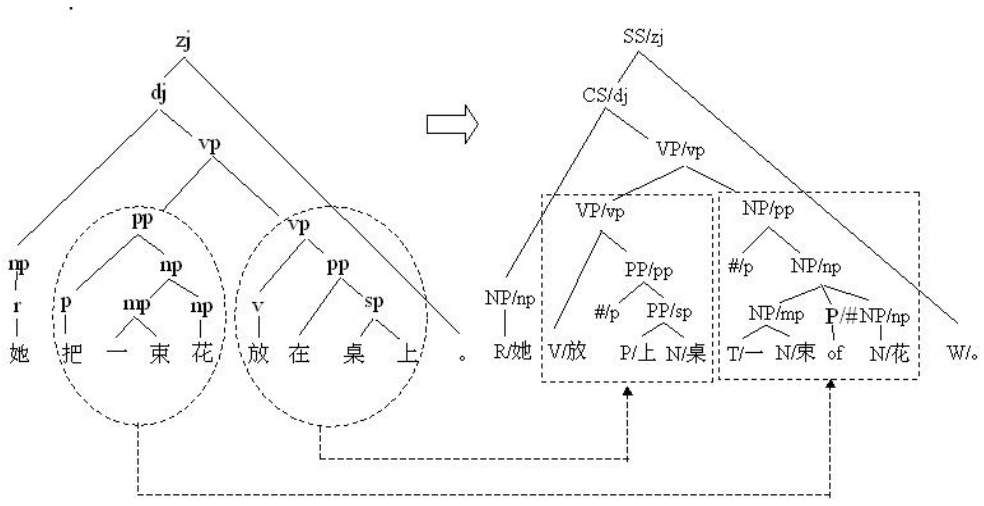


图-2

[3] 树形图（包括下面的图 2、3）中汉语短语标记以小写字母表示，英语短语标记以大写字母表示。zj 表示整句，dj 表示小句，np 表示名词性短语，vp 表示动词性短语，pp 表示介词性短语，mp 表示数词性短语，sp 表示处所词性短语，SS 表示英语的整句，CS 表示小句，其他大写标记含义与小写标记相同。W 是标点。

（4）对得到的目标语言结构进行进一步的调整。通常经过结构转换得到的句法结构还保留有源语言结构的诸多痕迹，需要根据目标语言的句法知识对该结构进行调整，对图-2所示译文结构进行调整后可以得到如下面的图-3 所示的新的译文结构。

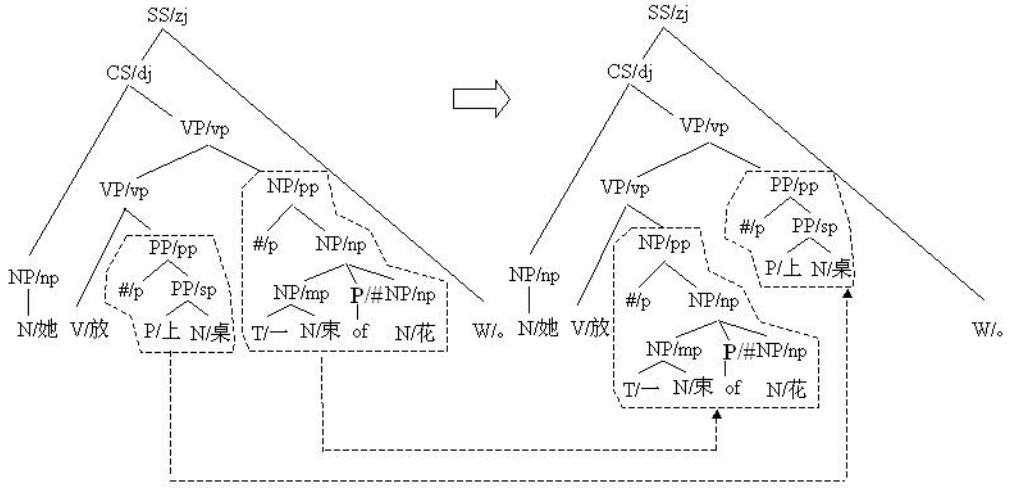


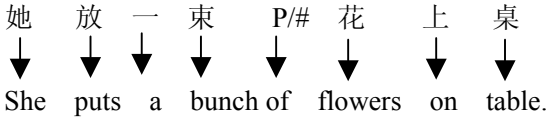
图-3

（5）对源语言文本中的单词进行译词选择，译词选择主要利用两种语言之间的词汇对译知识进行，译词选择并不是一件很容易完成的任务，通常一个源语言单词在目标语言中往往对应着多个单词（比如汉语的“开飞机，开门”都是“开”，但翻译成英文，分别要选择译词为“fly”和“open”），如何选择出正确的译词不仅要依赖于双语对照词典，同时还要综合考虑该单词的上下文环境。对于上述例子，存在下面的单词对应关系。

她↔she　　放↔put　　一↔a　　束↔bunch

花↔flower　　上↔on　　桌↔table　　。↔.

（6）利用得到的目标语言句法结构以及经过译词选择得到的目标词语生成目标语言词串，即将图-3 所示的译文句法结构树的叶子（leaf）节点取出，顺序排列就得到译文。如果目标语言是有形态的语言，还要进行目标语言单词的形态生成，把单词的原始形式变成合适的变体形式，比如本例中“put”变为第三人称单数形式“puts”，“flower”要变为复数形式“flowers”等。对于上述例子，最终计算机会产生出如下的英语译文（#表示空形式，译文中“of”是凭“空”增加的）。



从以上的扼要介绍可以看出，机器翻译涉及到源语言分析，源语言到目标语言的转换，目标语言的生成等几个大环节[4]，每一个环节都需要语言知识来告诉计算机做出正确的选择。因此，获取能够为机器翻译提供支持的语言知识，是机器翻译系统能否取得成功的一个关键因素。语言研究若要面向机器翻译做出贡献，就必须了解机器翻译的需求，从而进行有针对性的探索。下面我们就来探讨机器翻译向语言研究提出了什么样的要求。限于篇幅，下文的讨论将集中在汉外机器翻译中的汉语分析环节。不涉及转换和译文生成环节的问题，不过应该认识到，像上面这一句译文，大致看看，还过得去。但它很可能并不是一句地道的英语。在这句英语中，table 前面应该有冠词，put 是否要变成 puts 也要看上下文环境。因此，译文

[4] 基于规则的机器翻译系统也不一定都遵循本文描述的流程。而且不同系统在技术细节上还有许多差异。此外，基于统计的机器翻译系统和基于实例的机器翻译系统流程也都有自己的特点。可参看赵铁军（2000）

生成的研究也是十分重要而且艰巨的任务。尽管一个机器翻译系统各个环节涉及到的具体语言知识内容有差异，但下文对面向机器翻译的源语言分析环节的语言研究的认识也适用于转换和译文生成环节的语言研究。

二　机器翻译向语言研究提出了什么样的要求？

在以下讨论中，我们试图说明：跟面向人的语言研究一样，机器翻译也要求语言研究应该以发掘形式和意义之间的对应关系为目标。因为从语言研究的根本来说，无所谓面向人还是面向计算机，都是要对同形多义和同义多形的错综复杂的语言现象进行研究。不过，在具体操作层面上，面向机器和面向人的语言研究还是有差别的。因为机器和人对同形多义或同义多形，会有不同的界定标准，此外，机器跟人在具体的知识表示方面也存在差异，因而为适应机器翻译的需求开展语言研究，确实应该有一定的针对性。下面分三个方面展开讨论：

（一）机器翻译要求语言研究充分注意发掘形式与意义对应关系的重要性

自从 Chomsky 提出一个好的语言理论的三个标准是观察充分（observational adequacy），描写充分（descriptive adequacy），解释充分（explanatory adequacy）[5]以来，大多数语言学家已经把这三个充分作为语言理论所追求的目标[6]。中国的语法学者在进行语法研究时，强调研究的最高目标是揭示形式与意义之间的对应关系。这实际上可以看作是把上面“三个充分”贯穿起来的一个认识，即在观察语言现象，描写语言现象，解释语言现象等不同层面上，都要围绕一个共同的主题展开，就是清楚地说明一个语法形式跟它的语法意义之间的对应性。下面我们通过实例来简要说明以发掘形式与意义之间的对应关系为主旨的语法研究的一般模式。这可以从两个方面来看：

（1）从形式到意义：对表面上“相同（相似）”的形式进行分化，揭示其不同的意义。比如观察下面 A、B 两组例子之间的差别[7]：

A	B
在黑板上写字	在舞台上表演话剧
在墙上涂颜料	在车上看书
在郊区开饭馆	在会议室开会

A、B 两组例子的表面形式相同，都是“在 ＋NL＋V＋N”（其中 NL 表示处所成分，V 表示动词，N 表示名词）这样的格式（为称说方便，记作 S1）。但是，A 组例子表示的意思是动作完成后，N 所处的位置在 NL，而 B 组例子并不在意 N 的位置，它表达的是在 NL 这个处所发生了什么事件。格式 S1 对应的这两种不同意思可以通过变换方式显现出来，那就是 A 组例子都能变换成“把 ＋N＋V＋ 在 ＋NL”这个格式（记作 S2），而 B 组例子都不能做这样的变换：

A'	B'
把字写在黑板上	* 把话剧表演在舞台上
把颜料涂在墙上	* 把书看在车上
把饭馆开在郊区	* 把会开在会议室

---

[5] 参见 Chomsky（1965），之后随着研究的深化和细化，又有一些语言学家在此基础上做了发展，比如提出心理学充分，类型学成分，语用学充分等要求。

[6] 参见 Van Valin & Randy LaPolla（1997）所著 Syntax 一书。该书第一章将语言理论的目标表述为三个递进的方面：描述语言现象，解释语言现象，理解语言的认知基础，跟 Chomsky 的提法是一致的。

[7] 参见朱德熙（1978）。

不难看出，上面这样的研究模式，就是对“一个”多义形式进行分化。分化的对象可以是从词（多义词）到短语（多义短语）到句式（多义句式）等等大小不同的语言单位。

（2）从意义到形式：对表面上“意思相同”的形式进行辨析，揭示其在不同场合下的差异。比如观察下面甲、乙两组例子的异同[8]：

	甲	乙
A	1 去打球	1' 打球去
	2 去看电影	2' 看电影去
B	1 去寄钱	1' 寄钱去
	2 去坐火车	2' 坐火车去
	3 去请张三	3' 请张三去
C	1 *去让他	1' 让她去
	2 *去派他	2' 派他去

甲组例子是“去 ＋V＋N”格式（记作 S3），乙组例子是“V＋N＋ 去”格式（记作 S4）。仅看甲、乙两组里的 A 类例子，可以得出结论说，汉语中 S3 与 S4 可以表达“相同”的意思，都是在说动作行为的主体发生位移，而位移的目的是进行“V+N”所表示的活动。但是，随着观察范围的进一步扩大，我们就会看到 S3 和 S4 还有 B 类例子和 C 类例子这样的情况。B 类 S3 与 S4 也是都能成立，但表达的意思不完全一样，B 类甲组 S3 表达的意思是某人发生位移，位移的目的是进行“V+N”表达的动作行为，而 B 类乙组 S4 表达的意思则有各种不同的情况，B1'表示动作所支配的对象“钱”的位移趋向，B2'表示动作行为的主体以某种方式发生位移，B3'是递系结构，格式中的 N 主动发生位移；C 类 S3 不能成立，S4 可以成立。这样就显现出 S3 跟 S4 这两个格式之间的意思差异了。

同样不难看出，上面这样的研究模式，就是对“多个”同义（近义）形式进行辨析。发掘多个格式之间意思相同和相异的条件。

上述（1）（2）这两个方面又可以统一起来加以认识。从一个角度看，（2）是对表面上“意思相同”的两个形式（S3 和 S4）进行辨析，从另一个角度看，（2）又是对 S4 这一个“多义形式”进行分化（即 S4 可以解释为不同的语法意义）。因此无论是从形式到意义，还是从意义到形式的研究，遵循的实质上是同样的模式，追求的也是同样一个目标：那就是尽可能地去系统地整理语言中一个形式与另一个形式之间的对应关系。从计算机的角度看，这一点更为明显，对计算机来说，所谓“形式与意义之间的对应关系”，实质上等价于“一个形式与另一个形式之间的对应关系”（上文已经说过，机器翻译的过程就是被看作为一个在语言学知识引导下的符号变换过程）。

很显然，上述以揭示同形格式的细微差别为追求的语言研究模式，对机器翻译来说，无疑也是非常必要的。在这样的研究模式下产生的成果，有很多都可以直接转变为计算机可读的形式为机器翻译系统所利用[9]。比如上面建立在格式 S1，S2，S3，S4 等的异同比较基础上得到的语言知识，很多都可以在句法分析中发挥作用。换句话说，面向人开展的语言研究所积累的语言知识成果，同样可以为机器翻译提供帮助。当然，并不是直接就可以用，而是要针对机器的特点做相应的适应性调整。下面的讨论将说明：在具体操作层面上，机器翻译确实向语言研究提出了一些进一步的要求。

---

[8] 参看陆俭明（1985）

[9] 参见俞士汶（1999），詹卫东（1997）（2000）。

（二）机器翻译要求语言研究者所关注的语言现象的面要拓宽

自然语言中，形式与形式之间的对应关系错综复杂，常常会有一对多，多对一，甚至多对多这样的情况，不像交通信号灯组成的符号系统（红灯停、绿灯行、黄灯等），形式跟“意义”（形式）之间有明确的一一对应关系。而从机器的角度看形式与形式之间的对应关系，比从人的角度来看有更多的问题（可能在人看来很多问题都不成为“问题”）。下面我们分词处理层次上的问题和短语结构分析层次上的问题两方面来说明。

先看词处理层次上的问题。在词语处理的层次上，主要是如何在“字串形式”跟“词串形式”之间建立起对应关系，并进而在词串和词性标记串之间建立起对应关系的问题。比如本文第一节所举的例子。在翻译的第一个环节，就是将字串形式变换为词串及其词性标记串形式。

例 1

字串	她把一束花放在桌上。
词串	她　把　一　束　花　放　在　桌　上　。
词性标记串	r　p　m　q　n　v　p　ng　f　w

很显然，从字串到词串的形式变化过程，实际上是增加了信息（减少了不确定性）。对人来说，这个形式变换过程非常自然和容易。但对机器来说，在这个过程中，常常会碰到一对多的情形（人常常不会觉察到有“一对多”的问题），比如：

例 2

字串	明年开始地铁中将可以使用移动电话。
词串 1	明年　开始　地铁　中　将　可以　使用　移动　电话
词串 2	明年　开始　地铁　中将　可以　使用　移动　电话

人能很容易地将例 2 中的字串变换为词串 1 形式，但计算机却会面临在词串 1 和词串 2 之间进行选择的问题。其中字串“中将”[10]可能是两个词，也可能是一个词。人们一般把这个问题称为计算机分词中的组合歧义（参见刘开瑛 2000）。再看下面的例 3：

例 3

字串	张店区大学生不看重大城市户口
词串 1	张店区　大学生　不　看重　大　城市　户口
词串 2	张店区　大学生　不　看　重大　城市　户口

同样，计算机也会面临词串 1 和词串 2 之间的选择问题。其中字串“看重大”可能是“看重”跟“大”这两个词，也可能是“看”跟“重大”这两个词。人们一般把这个问题称为计算机分词中的交叉歧义问题。下面再看从词串到词性标记串的变换中会碰到的问题。

例 4

字串	把这篇报道编辑一下
词串	把　这　篇　报道　编辑　一　下
词性标记串 1	p　r　q　n　n　m　v
词性标记串 2	p　r　q　n　n　m　f
……	……

[10] “中将”作为一个词时其中的“将”读去声调，作为两个词时“将”读阴平调。从传统语言学的角度看，这里面根本就是两个不相干的“将”，一般不会把这里的“中将”联系在一起作为一个歧义问题看待。

例 4 是将词串变换为词性标记串时出现一对多的情形，按照北大计算语言所“现代汉语语法信息词典”（参看俞士汶等 1998）的词类划分，“把”有介词，动词，量词，名词四个词性标记，“这”有一个词性标记（代词），“篇”有一个标记（量词），“报道”和“编辑”都有名词和动词两个标记，“一”有数词和连词（c）两个标记，“下”有动词，方位词，量词三个标记，这样，这个词串对应的词性标记串就有 4×1×1×2×2×2×3=96 种可能性。人有能力在看到这个字串后马上把它变成正确的词串，进而在 96 种可能性中选择正确的一种词性标记串，计算机却不容易做到这一点。

要让计算机能准确地进行形式变换，就要告诉计算机相应的语言知识。知识可以用规则的形式表示，也可以用统计数据的形式表示。比如对于例 1，“上”的词性标记有可能是 v, q, f，“上”在例 1 所处的上下文环境是“在/p　桌/n + 上 + 。/w”，我们可以专门为“上”写一条规则，说明“上”如果左邻两个词为介词和名词，右邻句尾标记，那么“上”的词性应该标为方位词。但很明显，这样的规则形式有时候也会碰到问题，比如例 4 中如果用规则来判断“报道”的词性，可能会猜测左邻词为量词时，判定“报道”为名词，在例 4 的情形中，这样判断恰好是正确的，但在下面这个例句中，“报道”应该是动词：

例 5　这篇报道谢廷风的新闻是假的。

这里“报道”的左邻词环境跟例 4 的一样，但词性标记却不同。如果用上面的规则做判断依据，那么例 5 的词性标注就会出错。

针对规则方法的这类问题，人们又探索了用统计方法来进行分词和词性标注的处理。从目前的实践来看，在经过分词和词性标注处理的大规模语料库基础上训练得到的汉语分词和词性标注软件，达到了很好的处理效果[11]。限于篇幅，这里就不展开讨论了。

下面再看短语结构分析的层面的问题。在短语结构分析中，也有许多形式变换问题需要引起语言研究者的注意。这些问题可以统称为短语结构分析中的歧义问题。我们曾将歧义分为真歧义、伪歧义、准歧义等不同类型[12]，对人来说，比较容易注意到真歧义类型的歧义问题，但对计算机来说，后两种歧义类型也需要关注。从处理上说，伪歧义比较容易对付。比如计算机对第一节的例子“她把一束花放在桌子上”进行分析，就会碰到伪歧义格式的分析问题。“把一束花放在桌上”对应着“pp vp pp”这样的标记形式序列：

把一束花	放	在桌上
pp	vp	pp

对计算机来说，在短语标记层面，就发生一个问题，vp 到底是先左结合，还是先右结合？因为分析规则集中有这样两条动词短语规则，使得上面这个短语类序列在组合时碰到多选问题：　vp -> pp vp

vp-> vp pp

很显然，这两种结合方式并不造成更高层结合上的差异，因此，我们可以规定其中一种是正确的形式，避免产生不必要的分析结果。比如规定上面这个短语序列取 [pp [vp pp]]这样的结构形式。

下面再看一个准歧义的例子。

例 6　　A.　把衣物洗干净的方法　　B. 把群众检举的贪官

例 6A 跟 6B 都对应着“p<把> np vp 的 np”这个短语格式，但两个短语的结构分析结

[11] 北京大学计算语言所开发了人民日报标注语料库。中科院计算所利用 1 个月标注语料（200 万字）进行训练开发的 ICTCLAS 分词与词性标注系统在 973 评测中取得了很好的成绩，分词正确率最高达到 98.44%（法制领域文本），词性标注正确率最高达到 88.55%（国际新闻领域文本）。

[12] 参见詹卫东等（1999）。

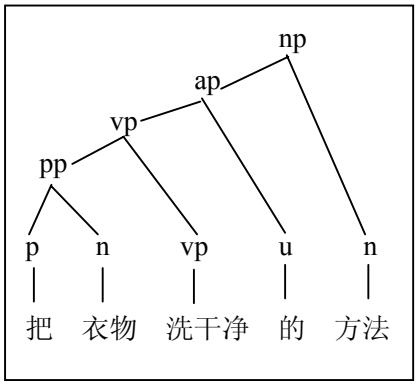
果却不同，6A 应该被分析为（或者说是将词串形式变换为句法结构串形式）：

[[[[把 衣物] 洗干净] 的] 方法]

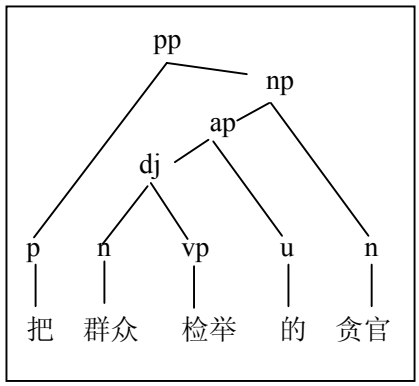
6B 应该被分析为：

[把 [[[群众 检举] 的] 贪官]]

在计算机中常以这种加括号的一维线性串方式来表达二维的树结构，人则更习惯看下面这样的图形表示方式：



(a)



(b)

对人来说，从例 6A 到树形图（a），从例 6B 到树形图（b），不存在形式转换的困难。但对计算机来说，如何做到不张冠李戴，把 6A 分析为 b，把 6B 分析为 a 呢？这就需要研究者针对计算机分析的特点来发现有用的语言知识了。像这样的问题，传统的面向人的语言研究一般不太关注。

上面分词汇处理和短语结构分析两个层面说明了机器翻译要求语言研究者关注的问题范围更宽一些。下面的例子可以进一步显示，汉语中有些形式对应的问题介乎词汇与短语之间。很难在所谓的词汇分析与短语分析之间划出一道截然的界限来。请看例子：

例 7　A　被侵略者逼上了绝路

B　被检举人走进了法庭

C　这个案子最终还是被调查人员找到了证据

D　80%的被调查学生赞同校长的处理方式

例 7 中划线部分都是“p<被>＋v＋n”形式的成分。例 7A 中“被侵略者”形成一个 pp 成分；7B 中“被检举人”更像是一个复合词；7C 中“被调查人员”也形成一个 pp 成分；7D 中“被调查学生”功能上相当于“被检举人”，但词感程度似乎低于“被检举人”的成词度。不管是词还是短语，这些例子都对应着“p<被>＋v＋n”这个形式，而有不同的分析结果，也是形式转换中一对多的问题。同样属于面向计算机的语言研究应该关注的现象。

上面这个例子是说不同层次的语言单位之间有界限模糊的情况，下面这个例子则试图说明，计算机处理自然语言的句子虽然分成多个环节，但应该充分注意到，各个环节之间是相互影响的，因此还要特别注意不同层次上的处理之间的相互关联性。比如：

例 8　有三百多种树

计算机可能将例 8 的结构层次分析为下面两种可能性：

A　有 三百多 种 树

vp　vp

B　有　三百多 种 树

v　np

上面这个例子，对人来说不大容易注意到它的歧义，一般都会很自然地把例 8 解释为 B 这个层次结构所对应的意思。但计算机分析时就会碰到歧义，因为其中的“种”实际上是个

多音字，对应着不同的词，读上声的“种”是量词，读去声的“种”是动词。在词处理阶段有两种不同的形式转换结果，在短语结构分析阶段也对应着两种不同的结果。即便人察觉到例 8 可能有两种解释，不过分化这个歧义多半也会从口语和书面这个角度去谈，因为在口语中，两个“种”读音不同，这样也就是不同的语音形式了，所谓歧义自然就不存在了。但对计算机来说，处理书面文本时碰到这个例子，仍然会有歧义问题，而且是同时牵涉到词处理层次和短语结构分词层次的歧义问题。此外，这里只讨论了一种语言内部的歧义问题。研究机器翻译，还应当关注两种语言之间的歧义问题（参见俞士汶 1989），限于篇幅，就不展开讨论了。

（三）机器翻译特别强调语言研究的结果有可操作性

这个要求主要体现在两个方面：（1）语言研究得到的成果是关于人类自然语言的知识，面向人的语言研究通常用自然语言来表述，而面向计算机进行语言研究，强调研究在形式化的框架下进行，研究所得到的语言知识应以形式化的方式表示。（2）面向机器翻译的语言研究特别强调语言知识的系统性及对真实语料的覆盖能力，即语言知识要达到一定的规模。

下面先谈第一个方面。

本文第一节中已经提到，在基于规则的机器翻译系统的各个环节中需要用到的语言知识都需要以计算机可懂的方式配备给计算机程序。这里不妨以一个简单的例子来加以说明什么是计算机可懂的形式。

比如“一件衣服”是汉语中一个具体的短语形式。如果用自然语言来描述这个短语对应的抽象规则（R），至少应包括这样三方面的内容：

1）作为一个整体，这个表达式是一个名词性范畴的语言成分；

2）这个表达式是由属于数量范畴的语言成分加上属于名词性范畴的语言成分组成的；

3）要实现这样一个组合，其中的名词性成分应该是能受数量成分修饰的那类名词，而且名词跟量词之间还需要满足一定的搭配关系。

上述规则 R 的作用是很明显的。它可以解释“一件衣服”在汉语中为“合法的”表达式，而排斥“* 一件纸”、“* 两个心胸”这样“非法的”表达式。如果计算机掌握了规则 R，当碰到例 a“一件纸做的衣服”，例 b“两个心胸宽阔的人”这样的形式时，就能够做出正确的判断：对于例 a，“一件”是修饰“衣服”的，而不是就近修饰同样也属名词范畴的“纸”；对于例 b，“两个”是修饰“人”的，而不是就近修饰也属名词范畴的“心胸”。

关于“件”能跟“衣服”搭配，不能跟“纸”搭配，“心胸”根本就不能跟任何个体量词搭配（不单单是不能跟“个”搭配），等等诸如此类的语言知识，是可以事先在一个人们称之为“词典”的地方一一加以记录的。也就是说，语法规则加上词典中的语言知识，可以构成计算机进行推理判断的已知条件（基础）。

在上面这个简单的例子中，规则以及跟规则配合使用的词典知识，都是用自然语言进行表述的。人容易理解，但计算机却不容易看懂。要让计算机掌握上述规则和相关的词典知识，就需要表述为下面这样的词典知识和规则形式：

词典

```
一  [词性:m,数词子类:基数]
件  [词性:q,量词子类:个体,表数:数]
衣服  [词性:n,名词子类:na,数量名:是,个体量词:件|套,…]
心胸  [词性:n,名词子类:ne,数量名:否,…]
```

规则

```
{R1} np -> mp !np   ::  $.内部结构=定中,$.定语=%mp,$.中心语=%np,$.dingyu=否,…, ①
                        %np.数量名=是,…                                          ②
                        IF %mp.量词子类=个体 THEN %np.个体量词=%mp.原形 ENDIF,… ③
{R2} mp -> m  !q    ::  $.内部结构=定中,$.定语=%m,$.中心语=%q,$.dingyu=是,…,
{R3} np -> !n       ::  $.内部结构=单词
```

在词典中以特征结构（即“特征名：特征值”）的方式记录了关于词语的语言知识。比如对词语“件”，用特征结构“量词子类:个体”表示“件”是个体量词。不同的特征以逗号分隔开。一个特征可以有多个取值，这种析取型取值用“|”分隔开，比如“衣服”这个词有特征“个体量词:件|套”，就表示“衣服”的“个体量词”特征取值既可以是“件”，也可以是“套”。“数量名:是”表示“衣服”前面可以受数量词的修饰（有些名词的前面不能受数量词的修饰，比如“心胸、笔者”）。在词典中记录的关于词语的特征结构描述，是在规则中进行合一运算的基础。

在“np -> mp !np”这条产生式规则（R1）中，箭头左部的 np 代表名词短语（比如“一件衣服”），右部的“mp !np”表示左部 np 短语是由一个数量短语（mp）加上一个名词短语（np）组成的。这样，一条上下文无关文法的产生式规则实际上刻划了自然语言成分的一个组合模式。

产生式右部 np 前的“!”符号标记一个成分在一个组合式中是中心成分（head）。中心成分是个技术概念。其作用是将中心成分的语法语义属性特征（特征结构）跟整个结构的语法语义属性特征（特征结构）关联起来。

“::”是分隔符，它后面是各种类型的合一运算表达式。这里用“$”符号代表产生式箭头左边的非终结符（即一条规则的根节点）；用“%”标记箭头右边符号的顺序[13]，“%mp”就表示箭头右边第一次出现的 mp；用“.”号表示对特征的引用，“$.内部结构”就表示箭头左部的 np 范畴的“内部结构”特征。

上面规则中合一表达式①里的“$.内部结构=定中”就是一个最简单的合一等式，合一的结果是箭头左部 np 的“内部结构”这一特征的取值为“定中”（合一之前，np 的“内部结构”取值为空，即未知）。“$.定语=%mp”表示左部 np 的“定语”特征取值为右部 mp。“$.dingyu=否”表示左部 np 整体不能再充当定中结构的定语（注意跟规则 R2 中的合一表达式“$.dingyu=是”对比）。

合一表达式②要求右部 np 的“数量名”特征值为“是”。这可以看作是一个测试条件。这样像“衣服”这样的词就可以通过这条规则的测试，而“心胸”就不能通过测试。

合一表达式③还引进了程序设计语言中常用的条件控制句“IF ... THEN ... ENDIF”来表示特定条件下才使用的合一，意思是：如果 mp 是由个体量词形成的数量短语，就执行 THEN 后面的合一运算，即要求 mp 后面的 np 的“个体量词”特征跟 mp 的“原形”特征（即 mp 的量词本身）匹配，否则不执行 THEN 后面的合一运算。

上面这个简单的示例实际上勾勒了以产生式规则、特征结构及合一运算结合起来表示语言知识的基本图景。目前主流的形式语法体系（包括 LFG，GPSG，HPSG 等），尽管技术细节各有不同，但其基本表达手段无不是由这三个部分组成，而在具体的知识系统的组成形式上，一般都可以划分为两块：规则库＋词库。在规则库中以产生式规则描述一种自然语言的结构模式，以合一关系来对各个结构模式进行约束；在词库中则以特征结构记录该语言词汇的多种属性特征（包括语法属性，语义属性等）。规则与词典相结合，将人类自然语言知识整理

[13] 这个规则只有一个“%”的情况，体会不到顺序问题。碰到“np->np !np”这样的规则，就需要两个“%”来区分规则右部第二个 np（%%np）跟第一个 np（%np）了。

成形式化的知识库，供计算机分析自然语言使用。

作为一个实用的语言知识系统，必然就涉及到上文谈到的可操作性问题的第二个方面，就是要求在形式化表达框架下建立起的语言知识系统要上规模。这也是机器翻译向语言研究提出的一个要求。之所以这样要求，主要有两个理由：第一：这样的成果能够直接为包括机器翻译在内的各种自然语言信息处理应用系统的开发服务，更具实用价值；第二：在大规模实践中更能考验一个语言理论体系的效能，同时对发展和改进一个语言理论体系来说，也是十分必要的。

广义的语义知识系统既包括词库和规则库这样的专家知识系统，也包括以带标记的真实语料（annotated corpus）来体现语言知识的语料库。国外学术界和信息产业界从上个世纪七八十年代以来陆续发展出一批大规模机器可读（machine readable）语言知识系统。其中有广泛影响的，有代表性的语言知识工程如美国普林斯顿大学的 WordNet，加州大学的 FrameNet，宾州大学的句法树库（UPENN Treebank），英国剑桥大学等单位开发的综合语言知识库（ILD），Lancaster 大学 UCREL 语料库中心的 LOB 语料库、句法树库（Lancaster-Leeds Treebank），美国微软公司开发的 MindNet 等，都是值得我国学者参考和借鉴的。这些词库和语料库作为公开的共享资源，既为语言的理论研究，也为自然语言处理系统的应用开发提供支持，极大地推动了研究工作的进展。

我国学者从上个世纪八十年代后期开始，逐步认识到形式化语言知识的大规模资源建设的重要性[14]。比如北京大学计算语言所和北京大学中文系合作，以朱德熙先生的词组本位语法理论体系为指导，逐步将这一理论体系下有关汉语句法结构的具体语言知识特征化，落实到数万词语的语法属性描述上，形成了目前规模已达 7.3 万词的《现代汉语语法信息词典》，在国内外中文信息处理领域已经产生很大影响，在包括机器翻译和信息检索在内的一些自然语言处理系统中发挥着重要的作用。除词语语法信息的大规模知识库外，我国学者在词语语义信息的大规模知识库建设方面也做出了许多努力，其中突出的代表性成果是董振东先生开发的面向概念描述的，包含中英文双语词条的“知网”（HowNet）知识库[15]。在词语语义信息描述方面，北大计算语言所延续了与北大中文系在建设语法信息词典的合作，从上个世纪九十年代末开始，在配价理论的指导下，分阶段完成了一部近 5 万词规模的中英文对照的语义词典。目前还以 WordNet 词库为参照，开发一部与 WordNet 规模相当的中文概念词典（CCD）。在建设大规模句法、语义词库的基础上，大规模真实语料库的加工和建设也相应地开展起来，从 1999 年开始到 2001 年，北大计算语言所历时近三年完成了人民日报语料库的词语切分、词性标注和全部人工校对工作。形成了一个 2600 多万字的标注语料库，这一资源已经开始对中文信息处理的一些应用产生积极影响（参见上文附注 11）。

除应用价值外，大规模语言实践对语言学理论本身的检验意义也是不言而喻的[16]。语言学前辈学者在谈到治学经验的时候常常提醒人们注意要有系统观念，全局意识，因为语言理论问题常常是“牵一发而动全身”。如果说在理论研究中领悟这样的要求多少还有些抽象的话，那么在进行大规模的语言工程实践的过程中，就能深切体会到这个经验之谈的重要性和实际意义。比如在面向机器翻译进行自动句法分析的需求全面归纳汉语的短语结构规则时，要写关于“把”字结构的规则，毫不夸张的说，就会牵涉到汉语全部的句法结构。因为一条规则会通过各种途径跟一个规则集中的其他多条规则发生关联，“把”字句中会涉及到汉语短语结构中其他几乎所有的范畴，这就时时提醒我们注意系统性，不能孤立地看待某个局部的语法现象。当我们宣称自己在研究“把”字句时，实际上并不是在圈定研究范围，而更应

---

[14] 参见陈力为、袁琦（1995）。

[15] 可访问 http://www.keenage.com/ 查看“知网”的最新资料。

[16] 提出一套语义表述理论固然不易（比如“格语法”），而要在数万词的规模上实践一套理论（不是用几个漂亮例子展示一下理论的魅力）就更是要费工夫了。在这样规模的实践过程中，理论存在的问题才能比较清晰地暴露出来，从而引导我们去改进原来的理论。

该看作是选择了一个研究视角，也就是从“把”字句这个角度去思考跟汉语各种句法结构都相关的全局性的问题。从这点上说，在形式化的知识表示框架下开展研究工作，可以更有效地提醒我们结构之间的相互关联，同时也让我们时刻保持清醒的认识：关于语言结构，哪些东西真正说清楚了，哪些东西暂时还没有办法说清楚。在计算机的词典里，是没有“含糊”这个词的，因此，在大规模的语言工程实践中，我们可以把计算机当作一面镜子，看看我们所笃信的语言理论信条是果如其然，还是并非如此。在不断实践与反思的交互作用中，一方面积累了宝贵的计算机可用的语言知识资源，一方面又为发展语言学理论提供了材料。

三 结 语

著名的计算语言学家 John Nerbonne 在一篇文章中这样描述理论语言学与计算语言学的关系：“在语言学和计算语言学之间存在着理论任务的自然分野，大致说来，语言学的责任是描述语言，而计算语言学提供算法和用于计算的体系结构。基于这种观点，这两个理论领域因为其共同关注的对象——语言——而发生紧密关系”。要让计算机能“翻译”一个句子，首先需要“理解”一个句子，而理解一个句子，实际上要解决下面这两个核心问题：

（1）一个句子的结构和意义是什么（如何呈现/表示一个句子的结构和意义）？

（2）如何得到一个句子的结构和意义?

第一个问题是“What”的问题，这是理论语言学关心的问题（语言学家也关心跟“What”相关的“Why”的问题，即一个句子的意义为什么是这样的，而不是那样的）；第二个问题是“How”的问题，这是计算语言学关心的问题，是面向机器翻译的语言研究需要关心的问题。本文以汉外机器翻译中汉语句法分析环节的需求为例，阐明了服务于机器翻译的语言研究应该加强对形式变换的研究，应拓宽所关注的语言现象的范围，应该重视研究成果的可操作性。这些方面都可以看作是为了回答“How”这个问题在进行努力。

而在具体落实上述要求时，语言学工作者应该比以往任何时候都重视研究手段和方式的更新。在信息时代的今天，语料库越来越容易得到，计算机检索和统计工具也越来越便于研究者使用，关系型数据库的可视化程度也越来越高，互联网的普及更是极大地提高了信息传播的效率，这些都为我们的研究工作提供了便利和更有效的手段。作为信息时代的语言学研究者，应该将语言研究的平台建立在新科技的基础上。

参考文献：

常宝宝，张伟（1998）《机器翻译研究的现状和发展趋势》，载《术语标准化与信息技术》，1998 年第 2 期。

陈力为、袁琦（1995）主编《中文信息处理应用平台工程》，电子工业出版社。

董振东（1998）《语义关系的表达和知识系统的建造》，载《语言文字应用》1998 年第 3 期。

刘开瑛（2000）《中文文本自动分词和标注》，商务印书馆。

陆俭明（1985）《关于“去+vp”和“vp+去”句式》，载《语言教学与研究》1985 年第 4 期。

陆俭明（1993）《八十年代中国语法研究》，商务印书馆。

俞士汶（1989）《机器翻译导引》，载《中国计算机用户》1989 年第 9 期。

俞士汶（1989）《自然语言的歧义与机器翻译对策》，载《中文信息学报》1989 年第 2 期。

俞士汶（1999）《自然语言理解与语法研究》，载马庆株编《语法研究入门》，商务印书馆。

俞士汶、朱学锋、王惠、张芸芸著（1998）《现代汉语语法信息词典详解》，清华大学出版社。

俞士汶、段慧明、朱学锋、孙斌（2002）《北京大学现代汉语语料库基本加工规范》，载《中

文信息学报》，2002 年第 5，6 期。
詹卫东（1997）《面向自然语言处理的现代汉语词组本位语法体系》，载《语言文字应用》1997 年第 4 期。
詹卫东、常宝宝、俞士汶（1999），《汉语短语结构定界歧义类型分析及分布统计》，载《中文信息学报》1999 年第 3 期。
詹卫东（2000）《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社。
赵铁军 等编著（2000）《机器翻译原理》，哈尔滨工业大学出版社。
朱德熙（1985）《语法答问》，商务印书馆。
朱德熙（1978）《“在黑板上写字”及其相关句式》，载《语言教学与研究》1978 年第 3 期。

Chomsky, Noam（1965）Aspects of the theory of syntax, Cambridge, MA: MIT Press.
Van Valin, Robert D., Jr. & LaPolla J. Randy（1997）Syntax: Structure, meaning and function, Cambridge University Press.
Nerbonne, John（1996），Computational Semantics -- Linguistics and Processing, In Shalon Lappin, ed. The Handbook of Contemporary Semantic Theory, Oxford: Blackwell, 1996, Chapter 17.
Fellbaum, Christiane（1998）ed., WordNet : an electronic lexical database, Mass :MIT Press.
Fillmore,C.J（1982） Frame semantics, In Linguistics in the morning calm, The Linguistic Society of Korea ed. Hanshin Publishing Co. Seoul, pp111-137.
Baker, Collin F., et al.（1998）The Berkeley FrameNet Project, In Coling’98. pp86-90.
Richardson, Stephen D., et al.（1998）MindNet: acquiring and structuring semantic information from text, In Coling’98.pp1098-1102.
ILD：http://www.cogs.susx.ac.uk/lab/nlp/carroll/ild.html
UCREL：http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html
UPenn: http://www.cis.upenn.edu/~treebank/

Machine Translation And Language Study

Abstract: First of all, this paper illustrates conventional workflow of rule-based machine translation systems. And then the authors try to answer the following question: what should linguists do with language formalization and representation required by study of machine translation? The authors claim that (1) the conditions of transformation between various linguistic forms should be paid more attention and studied systematically; (2) linguists should change their view on language ambiguity, for what ambiguity means in MT-oriented language research has been proved to be very different with human-oriented language research; (3) the linguistic knowledge should be readable or tractable by computer, and the scale of a knowledge base could also play a key role for practical purpose and the development of a linguistic theory.

Keywords: Machine translation, Language study, Electronic dictionary, Phrase Structure Rule, Formalization, Linguistic knowledge base