

载《中文信息处理若干重要问题》，科学出版社，2003年

机器翻译的现状和问题

The Current Situation and Problems in Machine Translation

冯志伟

教育部语言文字应用研究所计算语言学研究室

1. 机器翻译的现状:

1.1 机器翻译的回顾

■ 机器翻译思想的萌芽关于用机器来进行语言翻译的想法，远在古希腊时代就有人提出过了。在17世纪，一些有识之士提出了采用机器词典来克服语言障碍的想法。笛卡儿（Descartes）和莱布尼兹（Leibniz）都试图在统一的数字代码的基础上来编写词典。在17世纪中叶，贝克（Cave Beck）、基尔施（Athanasius Kircher）和贝希尔（Johann Joachim Becher）等人都出版过这类的词典。由此开展了关于“普遍语言”的运动。维尔金斯（John Wilkins）在《关于真实符号和哲学语言的论文》（An Essay towards a Real Character and Philosophical Language, 1668）中提出的中介语（Interlingua）是这方面最著名的成果，这种中介语的设计试图将世界上所有的概念和实体都加以分类和编码，有规则地列出并描述所有的概念和实体，并根据它们各自的特点和性质，给予不同的记号和名称。本世纪三十年代之初，亚美尼亚裔的法国工程师阿尔楚尼（G.B. Artsouni）提出了用机器来进行语言翻译的想法，并在1933年7月22日获得了一项“翻译机”的专利，叫做“机械脑”（mechanical brain）。这种机械脑的存储装置可以容纳数千个字元，通过键盘后面的宽纸带，进行资料的检索。阿尔楚尼认为它可以应用来记录火车时刻表和银行的帐户，尤其适合于作机器词典。在宽纸带上面，每一行记录了源语言的一个词项以及这个词项在多种目标语言中的对应词项，在另外一条纸带上对应的每个词项处，记录着相应的代码，这些代码以打孔来表示。机械脑于1937年正式展出，引起了法国邮政、电信部门的兴趣。但是，由于不久爆发了第二次世界大战，阿尔楚尼的机械脑无法安装使用。1903年，古图拉特（Couturat）和洛（Leau）在《通用语言的历史》一书中指出，德国学者里格（W. Rieger）曾经提出过一种数字语法（Zifferngrammatik），这种语法加上词典的辅助，可以利用机械将一种语言翻译成其他多种语言，首次使用了“机器翻译”（德文是ein mechanisches Uebersetzen）这个术语。

■ 1933年，苏联发明家特洛扬斯基（П. П. ТРОЯНСКИЙ）设计了用机械方法把一种语言翻译为另一种语言的机器，并在同年9月5日登记了他的发明。1939年，特洛扬斯基在他的翻译机上增加了一个用“光元素”操作的存储装置；1941年5月，这部实验性的翻译机已经可以运作；1948年，他计划在此基础上研制一部“电子机械机”（electro-mchanical machine）。但是，由于当时苏联的科学家和语言学家对此反映十分冷淡，特洛扬斯基的翻译机没有得到支持，最后以失败告终了

■ 1946年，美国宾夕法尼亚大学的埃克特（J. P. Eckert）和莫希莱（J.W. Mauchly）设计并制造出了世界上第一台电子计算机ENIAC，在电子计算机问世的同一年，英国工程师布

斯 (A.D.Booth) 和美国洛克菲勒基金会副总裁韦弗 (W.Weaver) 在讨论电子计算机的应用范围时, 就提出了利用计算机进行语言自动翻译的想法。

■1947年3月6日, 布斯与韦弗在纽约的洛克菲勒中心会面, 韦弗提出, “如果将计算机用在非数值计算方面, 是比较有希望的”。在韦弗与布斯会面之前, 韦弗在1947年3月4日给控制论学者维纳(N. Wiener) 写信, 讨论了机器翻译的问题, 韦弗说: “我怀疑是否真的建造不出一部能够作翻译的计算机? 即使只能翻译科学性的文章 (在语义上问题较少), 或是翻译出来的结果不怎么优雅 (但能够理解), 对我而言都值得一试。”可是, 维纳在4月30日给韦弗的回信中写道: “老实说, 恐怕每一种语言的词汇, 范围都相当模糊; 而其中表示的感情和言外之意, 要以类似机器翻译的方法来处理, 恐怕不是很乐观的。”

■1949年, 韦弗发表了一份以《翻译》为题的备忘录, 正式提出了机器翻译问题。在这份备忘录中, Weaver 写道:

“I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.”

这段话中, Weaver 首先提出了用统计方法进行机器翻译的想法, 成为后来噪声信道理论的滥觞。

Weaver 还提出了中间语言的概念, 他说 “The contents of source language and target language are the same”.

这个《备忘录》促进了机器翻译研究的热潮。50年代初期在美国和欧洲都出现了一些机器翻译的研究机构。美国政府对机器翻译的拨款达 1,000,000 美元之多。

■ Georgetown 大学的机器翻译试验 (1954): :把大约 50 个俄文句子 (化学文献) 翻译成英文。

■ 但是, 尽管取得了这些成绩, 机器翻译也出现了很多问题, 这些问题今天仍然在困扰着我们。机器翻译译文质量的低劣引起了普遍的失望, 投资部门开始对机器翻译的可行性进行调查。

■ ALPAC 报告 (1966年): ALPAC 对机器翻译采取否定态度的主要论据是经济方面的, 它指出机器翻译的费用再加上译后编辑的开支比纯粹的人工翻译的成本高, 但是, 我认为 ALPAC 报告并没有从学术上否定机器翻译, 在 ALPAC 报告中, 提出了加强面向计算机的语言研究的必要性, 首次使用了 “computational Linguistics” (计算语言学) 这个术语。ALPAC 报告导致了机器翻译在经济资助上的困难, 但同时也激发了计算语言学的研究。

■ 早期机器翻译研究者的主要失误在于: 他们低估了自然语言计算机理解的极端复杂性, 忽视了 Weaver 关于用统计方法研究机器翻译的忠告。

■ 低潮期间的机器翻译研究

--GETA (Grenoble, France): : 在 Bernard Vauquois 教授领导下, 开发了基于配价语法的机器翻译系统, 研制了机器翻译专用软件 ARIANE, 推动了逻辑程序设计的研究。

--TAUM-METEO (University of Montreal, 从 1977 开始): : 研制了实用性机器翻译系统 (English-French,) TAUM, 翻译天气预报文献, 在技术上, TAUM 继承了 GETA 的机器翻译方法。

-- SYSTRAN: 这个系统在 Apollo-Soyuz 空间研究方面承担了机器翻译的任务, 立下汗马功劳, 后来被 EEC 正式采用, 作为重要的翻译工具。

机器翻译的发展经历了一个马鞍形的过程:

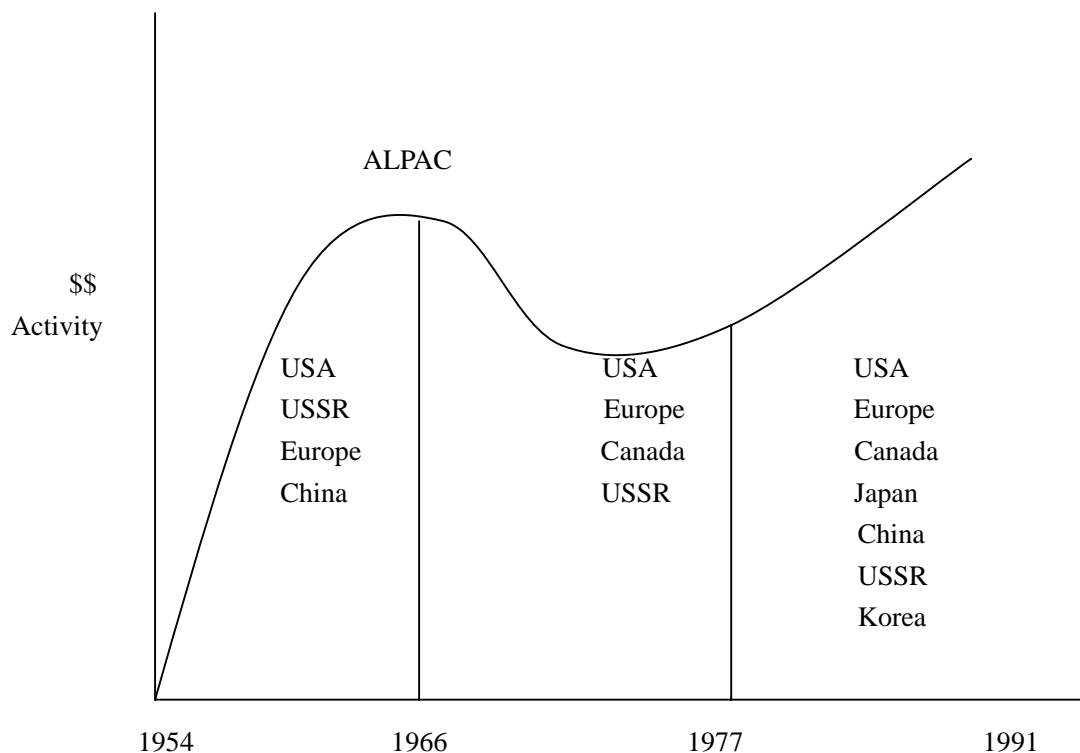


图 1 机器翻译发展的马鞍形过程

1.2 基于语言规则的机器翻译方法

■ 直接翻译系统：翻译系统的设计基本上是根据原语（source language, SL）和目标语（target language, TL）之间的词汇单元的对应关系，因此这样的系统带有针对性过强（ad hoc）的弊病。

例如，在英语-西班牙语的直接翻译系统中，英语句子“I like Mary”用直接翻译方法就单词对单词地翻译成“Me (I) gusta (like) Maria”，但是正确的西班牙语句子应该是“Maria me gusta”（“Mary to me pleases”）。在一般情况下，原语句子的词序和目标语句子的词序是不相同的，因此，如果用户对于目标语的结构没有足够的知识，很难理解用直接翻译系统翻译出来的句子。

有些直接翻译系统也考虑目标语的一些句法特性和词序规律，添加一些规则来改善目标语译文的可读性。例如。上面的英语-西班牙语翻译系统可以添加如下的直接映射规则：

$X \text{ LIKE } Y \rightarrow Y \text{ X GUSTAR}$

这样的规则可以明显地改善译文的可读性。不过，这种规则已经带有转换系统的特点了。

■ 转换系统：原语的分析独立于目标语，原语的分析一般只在句法平面上进行。转换时需要一部双语对应词典，用目标语单元替换原语单元时应该考虑上下文。转换时还要考虑原语和目标语的结构差别，进行结构转换。

下面是西班牙语到英语的一条转换映射规则

$\text{gustar (SUBJ(ARG2:NP), OBJ1(ARG1:CASE))} \rightarrow \text{like (SUBJ(ARG1:NP), OBJ1(ARG2:NP))}$

这条转换规则规定：当西班牙语 *gustar* 转换为英语的 *like* 时，主语和宾语的论元都要改变位置，西班牙语的 *OBJ1* 要变格，转换成英语时成为 *SUBJ*，而且没有格的变化。与直接翻译系统不同，转换系统的建造需要进行双语对比，构造复杂的映射规则。

■ 中间语言系统：原语与目标语不直接接触。把原语的文本用人工设计的无歧义的

中间语言（interlingua）表示出来，然后再把中间语言所表达的意义用目标语言的词汇和句法结构表示出来。

中间语言表示的例子：

Like/gustar: [CAUSE (X, [BE (Y, [PLEASED])])]

其含义是：“某物或某人(X)引起某人(Y)喜欢”

中间语言系统不需要转换规则，因为中间语言表达式对原语和目标语都是一样的。

目前设计的中间语言系统有两种类型：

--大多数面向人工智能的机器翻译系统：其中有的系统有叫做基于知识的系统（knowledge-based MT，简称 KBMT）。主要有： .

a. Magaret Masterman (Cambridge language Research Unit, Britain)的机器翻译系统； ,

b. 优选语义学 (Wilks): 中间语言是基于 80 个语义基元 (primitives)，用树结构或网络结构表示；

c. CD 理论 (Schank 及其学派): 基于 MARGIE 自然语言理解模型的英德机器翻译系统。

d. KBMT 机器翻译系统: Yale 大学的 KBMT (Jaime Carbonell, Richard Cullingford 和 Anatole Gershman); Colgate 大学的 KBMT (Sergei Nirenburg, Victor Raskin 和 Tucker) .

e. HICATS/JE 机器翻译系统: 使用语义基元 (semantic primitives) 来组织词典结构

-- 不基于知识的中间语言系统: 与具体的语言知识无关，主要设计一种中间语言的表达工具，如 CETA, DLT, Rosetta 等系统。

基于语言规则的机器翻译方法可总结为如下的机器翻译金字塔： .

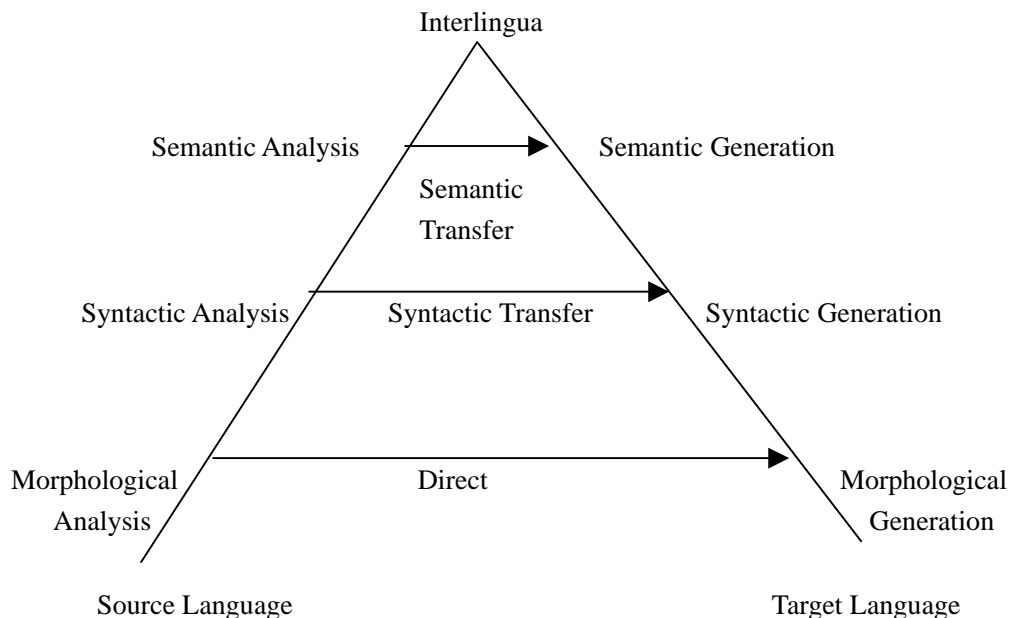


图 2 机器翻译金字塔

很多学者对于中间语言系统持怀疑态度。下面日本学者 Nagao 和德国学者 Schneider 的意见：

-- Makoto Nagao 说: :“.. when the pivot language [i.e. interlingua] is used, the results

of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility.”(见“Machine Translation”, Oxford, 1989).

-- Patel-Schneider 说: ”METAL employs a modified transfer approach rather than an interlingua. If a meta-language [an interlingua] were to be used for translation purposes, it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight.”(见“A four-valued semantics for terminological reasoning”, Artificial Intelligence, 38, 1989)

Nagao 认为中间语言不可能, Schneider 认为, 即使设计出中间语言的机器翻译系统, 它的管理会非常困难, 系统将由于不堪重负而自行崩溃。

我们认为, 在基于语言规则的机器翻译系统中, 比较可行还是转换系统。目前国际上比较好的基于规则的机器翻译系统, 大部分都是基于转换的系统。例如, SailLabs 公司的机器翻译系统就是基于转换的系统。

SailLabs 公司是一家专门从事翻译技术、内容技术和语音技术的德国公司, 其总部在德国的慕尼黑, 在德国的柏林和西班牙的巴塞罗那设有分公司。该公司的机器翻译系统前身是 Siemens (西门子) 公司开发的 METAL 系统。

该公司开发的机器翻译系统严格遵守“分析—转换—生成”的模式。分析分为 Morphological Analysis、Syntax Analysis、Mirification 三个过程, 其中前两个过程就是一般的形态分析和句法分析, 由于在该系统中, 句法分析按照语言学的理论来进行(如 X 阶标理论), 产生的句法树上有很多的空结点和功能性结点, 句法树的层次很深, 为了转换的经济性起见, 在转换之前, 需要将这棵句法树“压扁”, 这个过程称为 Mirification。转换发生在句法树层面。每一种不同的语言对的翻译系统都需要专门的语言学家开发相应的转换模块。

目前该公司已经有英法、英德、英西、德法等四种语对的双向机器翻译系统, 翻译正确率都比较高, 译文可读性强。

1.3 基于语料库的机器翻译方法

基于语料库的机器翻译方法有可分为两种: 基于统计的机器翻译方法和基于实例的机器翻译方法, 这两种都使用语料库作为翻译知识的来源, 所以可以统称为基于语料库的机器翻译方法。这两种方法的区别在于:

- 在基于统计的机器翻译方法中, 知识的表示是统计数据, 而不是语料库本身; 翻译知识的获取是在翻译之前完成, 翻译的过程中不再使用语料库。
- 在基于实例的机器翻译方法中, 双语语料库本身就是翻译知识的一种表现形式(不一定是唯一的), 翻译知识的获取在翻译之前没有全部完成, 在翻译的过程中还要查询并利用语料库。

早在 1947 年, 韦弗在他的备忘录中, 就提出了使用统计学的办法来解决机器翻译问题, 但是, 由于当时尚缺乏高性能的计算机和联机语料, 采用基于统计的机器翻译在技术上还不成熟。现在, 这种局面已经大大改变了, 计算机在速度和容量上多有了大幅度的提高, 也有了大量的联机语料可供统计使用, 因此, 在 90 年代, 基于统计的机器翻译又兴盛起来。在 Waever 思想的基础上, IBM 公司的 Brown 等人提出了统计机器翻译的数学模型。

基于统计的机器翻译把机器翻译问题看成是一个噪音信道问题, 如图所示:

$S \rightarrow \text{噪音信道} \rightarrow T$

可以这样来看机器翻译: 一种语言 S 由于经过了一个噪音信道而发生了扭曲变形, 在

信道的另一端呈现为另一种语言 **T**，翻译问题实际上就是如何根据观察到的语言 **T**，恢复最为可能的语言 **S**。语言 **S** 是信道意义上的输入，在翻译意义上就是目标语言，语言 **T** 是信道意义上的输出，在翻译意义上就是源语言。从这种观点看来，一种语言中的任何一个句子都有可能是另外一种语言中的某几个句子的译文，只是这些句子的可能性各不相同，机器翻译就是要找出其中可能性最大的句子，也就是对所有可能的目标语言 **S** 计算出概率最大的一个作为源语言 **T** 的译文。由于 **S** 的数量巨大，可以采用栈式搜索(stack search)的方法。栈式搜索的主要数据结构是表结构，表结构中存放着当前最有希望的对应于 **T** 的 **S**，算法不断循环，每次循环扩充一些最有希望的结果，直到表中包含一个得分明显高于其它结果的 **S** 时结束。栈式搜索不能保证得到最优的结果，它会导致错误的翻译，因而只是一种次优化算法。

基于统计的机器翻译进行概率计算时，采用隐马尔可夫模型(Hidden Markov Model, 简称 HMM)。隐马尔可夫模型是马尔可夫模型的扩展。马尔可夫模型描述的是一个随机过程，而隐马尔可夫模型中有两个随机过程，一个随机过程描述观察值(例如，具体的单词)和状态(例如，该单词可能标注的词类)之间的概率关系，即观察值是状态的概率函数，另一个随机过程描述状态之间(例如，词类标记与词类标记之间)的转移关系。作为外界的观察者来说，只能看到状态产生的观察值，而看不到状态之间的转移，状态之间的转移是隐藏的，所以叫做隐马尔可夫模型。近年来，利用隐马尔可夫模型在词性标注方面取得了较好的结果，从而推动了基于统计的机器翻译的研究。

比较著名的基于统计的机器翻译系统是 IBM 公司的 **Candide** 系统 [Berger 1994]。

IBM 公司 **Peter Brown** 等研究者基于统计机器翻译的思想，以英法双语对照加拿大议会辩论记录作为双语语料库，开发了一个英法机器翻译系统 **Candide**。

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		
Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

上表是 ARPA (美国国防部高级研究计划署) 对几个机器翻译系统的测试结果，其中第一行是著名的 **Systran** 系统的翻译结果，第二行是 **Candide** 的翻译结果，第三行是 **Candide** 加人工校对的结果，第四行是纯人工翻译的结果。评价指标有两个：**Fluency** (流利程度) 和 **Adequacy** (适当程度)。**Transman** 是 IBM 研制的一个译后编辑工具。**Time Ratio** 显示的是用 **Candide** 加 **Transman** 人工校对所用的时间和纯手工翻译所用的时间的比例。从指标上看，**Candide** 已经超越了采用传统方法的商品系统 **Systran**。

据报道，这个机器翻译系统包括三个部分：

- 英语的三元语法模型；
- 法语的三元语法模型；
- 英语和法语的部分对齐句子的高质量的对应该模型。

由于计算的复杂性，**Candide** 请了一些语言学家来帮助他们做形态分析表、语义标注、中间表达式的转换，**Candide** 也使用了词典。可见，这个系统还不能说是纯统计的。

事实上，纯统计的机器翻译的思想是注定要失败的。在这个系统中，纯统计方法所得到的机器翻译结果，其句子的正确率只有 40%，正是由于补充了上述的语言学规则，才使 **Candide** 系统得到了上述的结果。

IBM 的这个统计机器翻译系统后来由于外部和内部的财政支持都撤走了，因此，这个

系统的工作只坚持到 1995 年。

可见，统计方法是令人鼓舞的，可是它并不能解决所有困难的问题。

美国著名机器翻译学者 Yorick Wilks 在批评 Candide 系统时指出：“他们在系统中引入符号结构就说明了，纯统计的假设已经失败了”（“Incorporating symbolic structures shows the pure statistics hypothesis has failed.”）。

统计机器翻译系统，除了基于噪声信道理论的系统之外，还有基于最大熵方法的系统。

A.L.Berger 在 1996 年提出自然语言处理中“最大熵方法”(Maximum Entropy Approach)。德国 Aachen 技术大学的 Och 等发现，把 IBM 统计机器翻译基本方程式中的翻译模型换成反向翻译模型，总体的翻译正确率并没有降低，因此，他们提出基于最大熵方法的机器翻译模型，这样的系统使用了最大熵原理。所谓“最大熵原理”，就是说：对于一个随机事件，如果我们已经建立了一组样例，我们希望再建立一个统计模型，来模拟这个随机事件的分布。为此，我们需要选择一组特征，使得我们得到的这个统计模型，在这一组特征上，尽可能地与样例中的分布一致，同时又保证这个模型尽可能地均匀。以保证除了这一组特征之外，这个模型没有其他的任何偏好，也就是使模型的熵值达到最大。根据最大熵原理，在只掌握部分样本信息的情况下，应当选取符合已知知识并且具有最大熵的统计推断，也就是选择当前样本数据制约下的最不确定分布，所有已知的知识以特征形式出现。最大熵模型的优点表现在建模时只需要集中精力选择合适的特征集合，模型以参数的形式体现这些特征对模型的贡献，特征的选取不受历史条件的制约，可以灵活进行；对于数据稀疏问题，最大熵模型也不需要专门的平滑算法，而可以通过特征选择来进行数据平滑。应用最大熵方法，Och 等首先将噪声信道模型中的翻译模型转换成反相的翻译模型，简化了搜索算法，但翻译系统的性能并没有下降。然后他们又引入了一些参数和特征来调制，使系统的性能更加提高。Och 关于最大熵机器翻译的论文，得到 ACL2002 的最佳论文奖。

基于实例的机器翻译 (Example-based MT，简称 EBMT) 的思想最早是由日本机器翻译专家长尾真(Makoto Nagao)提出来的。他在 1984 年发表了《采用类比原则进行日-英机器翻译的一个框架》一文，探讨日本人初学英语时翻译句子的基本过程，长尾真认为，初学英语的日本人总是记住一些最基本的英语句子以及一些相对应的日语句子，他们要对比不同的英语句子和相对应的日语句子，并由此推论出句子的结构。参照这个学习过程，在机器翻译中，如果我们给出一些英语句子的实例以及相对应的日语句子，机器翻译系统来识别和比较这些实例及其译文的相似之处和相差之处，从而挑选出正确的译文。长尾真指出，人类并不通过做深层的语言学分析来进行翻译，人类的翻译过程是：首先把输入的句子正确地分解为一些短语碎片，接着把这些短语碎片翻译成其它语言的短语碎片，最后再把这些短语碎片构成完整的句子，每个短语碎片的翻译是通过类比的原则来实现的，也就是“通过类比来进行翻译”（“translation by analogy”）。因此，我们应该在计算机中存储一些实例，并建立由给定的句子找寻类似例句的机制，这是一种由实例引导推理的机器翻译方法，也就是基于实例的机器翻译方法。

在基于实例的机器翻译系统中，系统的主要知识源是双语对照的翻译实例库，实例库主要有两个字段，一个字段保存源语言句子，另一个字段保存与之对应的译文，每输入一个源语言的句子时，系统把这个句子同实例库中的源语言句子字段进行比较，找出与这个句子最为相似的句子，并模拟与这个句子相对应的译文，最后输出译文。

基于实例的机器翻译系统中，翻译知识以实例和义类词典的形式来表示，易于增加或删除，系统的维护简单易行，如果利用了较大的翻译实例库并进行精确的对比，有可能产生高质量译文，而且避免了基于规则的那些传统的机器翻译方法必须进行深层语言学分析的难点。在翻译策略上是很具有吸引力的。

要进行基于实例的机器翻译需要研究如下问题：

第一，正确地进行双语自动对齐(alignment)：在实例库中要能准确地由源语言例句找到相应的目标语言例句，在基于实例的机器翻译系统的具体实现中，不仅要求句子一级的对齐，而且还要求词汇一级甚至短语一级的对齐。

第二，建立有效的实例匹配检索机制：很多研究者认为，基于实例的机器翻译的潜力在于充分利用短语一级的实例碎片，也就是在短语一级进行对齐，但是，利用的实例碎片越小，碎片的边界越难于确定，歧义情况越多，从而导致翻译质量的下降，为此，要建立一套相似度准则(similarity metric)，以便确定两个句子或者短语碎片是否相似。

第三，根据检索到的实例生成与源语言句子相对应的译文：由于基于实例的机器翻译对源语言的分析比较粗，生成译文时往往缺乏必要的信息，为了提高译文生成的质量，可以考虑把基于实例的机器翻译与传统的基于规则的机器翻译方法结合起来，对源语言也进行一定深度的分析。

目前世界上的基于实例的机器翻译系统主要有：

-- 日本京都大学长尾真和佐藤(S. Sato)的 MBT1 和 MBT2 系统：MBT1 只能利用句子的格框架来选择适当的译文，实际上只是一个基于实例的译文选择系统。MBT2 是一个完整的基于实例的机器翻译系统，该系统的翻译过程分为分解(decomposition)、转换(transfer)、合成(composition)三步。在分解阶段，系统根据提交的源语言词汇依存树检索实例库，并利用检索到的实例碎片来表示该源语言句子的依存树，形成源匹配表达式；在转换阶段，系统利用实例库中的对齐信息将源匹配表达式转换成目标匹配表达式；在合成阶段，将目标匹配表达式展开成为目标语言词汇依存树，输出译文。

-- 美国卡内基-梅隆大学的多引擎机器翻译系统(Multi-engine Machine Translation) PANGLOSS 系统：这个系统的主要引擎是基于知识的机器翻译系统，基于实例的机器翻译系统只是它的一个引擎，为整个多引擎机器系统提供候选结果。下面我们还要进一步介绍这个多引擎机器翻译系统。

--日本口语翻译通信研究实验室 ATR 的 ETOC 和 EBMT 系统：ETOC 系统能够检索出与给定的源语言句子相似的实例，EBMT 系统能够利用实例库来消解歧义，这两个基于实例的机器翻译系统还不完整。

我国清华大学计算机系也进行了基于实例的机器翻译试验，建立了基于实例的日汉机器翻译系统；在哈尔滨工业大学和清华大学联合开发的计算机写作和翻译的集成环境“达雅”系统中，也使用了基于实例的技术。

1.4 多引擎机器翻译

Frederking 提出了一种典型的多引擎机器翻译 (Multi-Engine MT) 的方法 (1994)。该方法基本思想描述如下：

- 多个的翻译引擎同时对输入的句子进行翻译，不仅仅对整个句子进行翻译，而且对句子的任何一个片断也可以给出相应的译文，同时对这些译文片断给出一个评分。
- 各个翻译引擎共享一个类似线图 (Chart) 的数据结构，根据原文片断所处的位置，将它们的译文片断放在这个公共的线图结构之中。
- 对各个引擎给出的片断的评分进行一致化处理，使之具有可比较性。
- 采用一个动态规划算法 (称为 Chart Walk 算法)，选择一组恰好能够覆盖整个原文输入句子，同时又具有最高评分的译文片断，作为最后输出的译文。

多引擎机器翻译系统结构如下图所示：

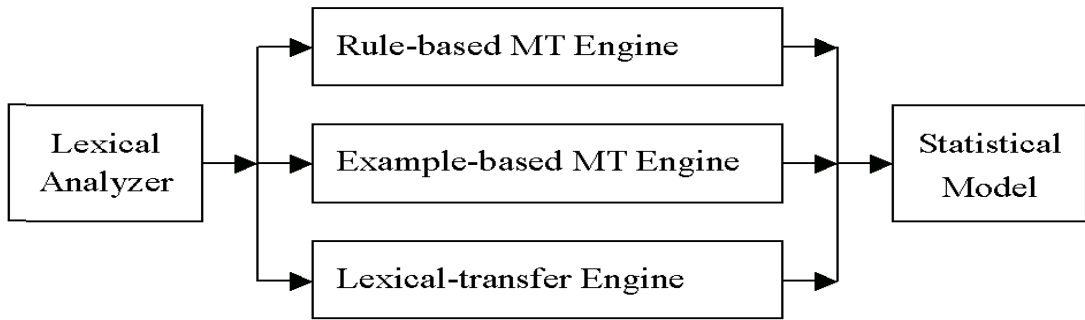


图3 多引擎机器翻译的系统结构

Hogan 通过一个简单的实验，证明这种方法确实可以得到比任何一种单一的方法都更高的准确率（1998）。

在多引擎奇迹翻译系统中，翻译的知识表示和知识处理是非常重要的，所以，多引擎机器翻译系统应该是一个基于知识的机器翻译系统。

美国卡内基-梅隆大学机器翻译中心研制的多引擎西班牙—英语机器翻译系统 PANGLOSS 系统就是一个这样的基于知识（Knowledge-Based）的机器翻译系统（KBMT）。这个系统充分利用语言中蕴藏的丰富的形态、句法和语义知识，他们还开发了建立本体知识体系（ontology）的工具，提出了一个自动获取词汇知识的框架。这个 KBMT 的重点是研制低层知识表示（Knowledge representation, 简称 KR）的方法。

KBMT 的知识源主要有：

- 形态变化表（Morphology tables）
- 语法规则（Grammar rules）
- 词表（Lexicons）
- 世界知识的表示（Representation of world knowledge）

PANGLOSS 系统总共包括三个翻译引擎：一个基于转换的翻译引擎（Transfer MT）、一个基于知识（KBMT）的翻译引擎和一个基于实例的翻译引擎（EBMT）。其系统结构如下图所示：

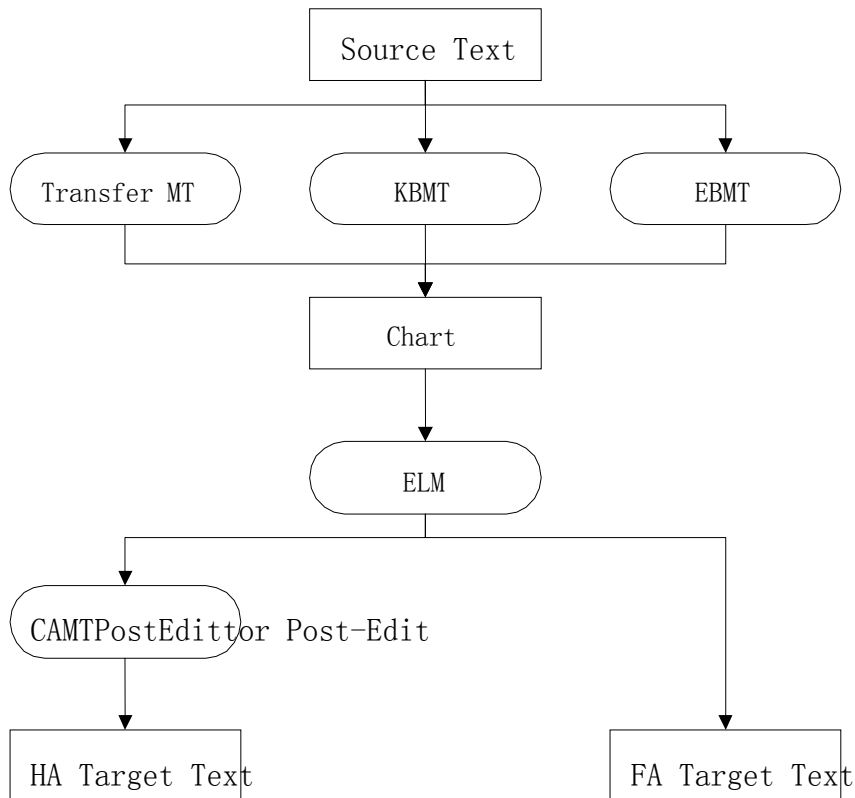


图 4 Panglos 机器翻译系统的结构

其中 CAMT PostEdit 是计算机辅助译后编辑(Computed-Aided Post-Edit)。HA Target Text 指人工辅助生成的目标语文本 (Human-Aided Target Text), FA Target Text 指全自动产生的目标语文本(Fully-Automated Target Text)。ELM 是英语语言模型 (English Language Model)。

在 EBMT 翻译引擎中,不需要任何的语言结构知识,需要的语言资源是一个双语例句库,一部双语电子词典和一个目标语言的同义词分类器。其中的双语例句库的规模达到了 72 万西班牙—英语语句对,主要来自于 UN 多语言语料库 (UN Multilingual Corpus), 语言数据联盟 (Linguistic Data Consortium); 而目标语言的同义词分类器是从词网 (WordNet) 中提取得到的,连同双语词典一起来寻找原语和目标语的语句中词语的关联。据统计,该 EBMT 引擎对于不限制领域的输入能够达到 70.2%的覆盖率。

实现方法包括两个步骤:首先通过查找例句库的索引寻找同输入匹配的最长组块 (chunk), 然后进行语句片段的对齐,确定组块的译文。

为了能有效利用已有的例句,他们提出了对例句进行了泛化 (Generalization) 的思想。

例如下面的英语和德语:

John Hancock was in Philadelphia on July 4th.

John Hancock war am 4. Juli in Philadelphia.

如果能知道 “John Hancock” 是一个人, “Philadelphia” 是一个城市, 而 “July 4th” 是一个日期的话, 就可以将上面的语句对简化为:

<PERSON> was in <CITY> on <DATE>.

<PERSON> war am <DATE> in <CITY>.

其中的<PERSON>, <CITY>和<DATE>代表了一些特殊的词语类。系统通过查找它的知识库来判断 “John Hancock” 是一个人等。当例句输入到例句库的时候,系统通过一系列的查找和替换将类似的词语替换成为这些特殊的“符号”。被替换的词语和其对应的翻译仍然被保存,用于将来的翻译。

在 PANGLOSS 系统中除了 EBMT 翻译引擎外,还有一个转换引擎和基于知识的翻译引擎 KBMT。对于一个输入语句,每一个翻译引擎都可能对其片段产生译文,然后把这些产生的译文放到一个统一的线图中,最后根据一个英语的统计语言模型 (ELM) 来决定线图 (Chart) 中的最佳路径作为译文输出,这样做可以尽量结合各个翻译引擎的优点,有利于产生最好质量的译文。与前述 Frederking 提出的 Chart Walking 算法的区别在于,这里用于选择最佳路径的函数不是根据各个翻译引擎给出的评分,而是根据英语的统计语言模型来决定,哪一个译文在英语中出现的可能性最大。

这个多引擎机器翻译系统的经验是:

- 根据使用目标的不同选择不同的翻译引擎
- 基于统计的机器翻译引擎最适合于翻译科技文摘。

1.5 规则方法和统计方法相结合

在自然语言处理中关于规则方法和统计方法的反映了语言学中的理性主义思潮与经验主义思潮的对立。有一些学者往往持相当极端的观点。Noam Chomsky 早在 1956 年就说, “But it must be recognized that the notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.” (“然而应当认识到, ‘句子的概率’ 这个概念, 在任何已知的对于这个术语的解释中, 都是一个完全无用的概念。”) Chomsky 完全无视“句子的概率”, 对于统计方法嗤之以鼻。

而 IBM 公司 Watson 研究中心语音研究组的负责人 Fred Jelinek 在 1998 年 (当时他在

IBM 语音研究组)更是语出惊人,他说:“Anytime a linguist leaves the group the recognition rate goes up” (“语言学家离开我们的研究组, 语音识别率就提高一步¹。”)应该说, 这两位著名学者的意见都是很偏颇的。

更多的学者在探索把规则方法和统计方法相互结合的途径, 并且取得了很好的成果。他们的研究主要包括两方面, 一是提出概率上下文无关语法, 二是提出词汇化的概率上下文无关语法。

■ 概率上下文无关语法 (Probabilistic Context-Free grammar, 简称 PCFG)

概率上下文无关语法又叫做随机上下文无关语法 (Stochastic Context-Free Grammar, 简称 SCFG)。

上下文无关语法 G 可以定义为四元组 $\{N, \Sigma, P, S\}$ 。而 PCFG 则在每一个规则 $A \rightarrow \beta$ 上增加一个条件概率 p :

$$A \rightarrow \beta [p]$$

这样, PCFG 就可定义为一个五元组 $G = \{N, \Sigma, P, S, D\}$, 其中 D 是给每一个规则指派概率 p 的函数。这个函数表示对于某个非终极符号 A 重写为符号串 β 时的概率 p 。这个规则可写为:

$$P(A \rightarrow \beta)$$

或者写为:

$$P(A \rightarrow \beta | A)$$

从一个非终极符号 A 重写为 β 时应该考虑一切可能的情况, 并且其概率之和应该等于 1。

例如, 我们有如下的 PCFG 规则:

- $S \rightarrow NP VP$ [0.80]
- $S \rightarrow AUX NP VP$ [0.15]
- $S \rightarrow VP$ [0.05]
- $NP \rightarrow Det Nominal$ [0.20]
- $NP \rightarrow Proper Noun$ [0.35]
- $NP \rightarrow Nominal$ [0.05]
- $NP \rightarrow Pronoun$ [0.40]
- $Nominal \rightarrow Noun$ [0.75]
- $Nominal \rightarrow Noun Nominal$ [0.20]
- $Nominal \rightarrow Proper-Noun Nominal$ [0.05]
- $VP \rightarrow Verb$ [0.55]
- $VP \rightarrow Verb NP$ [0.40]
- $VP \rightarrow Verb NP NP$ [0.05]
- $Det \rightarrow that$ [0.05] | $this$ [0.80] | a [0.15]
- $Noun \rightarrow book$ [0.10] | $flight$ [0.50] | $meat$ [0.40]
- $Verb \rightarrow book$ [0.30] | $include$ [0.30] | $want$ [0.40]
- $Aux \rightarrow can$ [0.40] | $does$ [0.30] | do [0.30]
- $Proper Noun \rightarrow Houston$ [0.10] | $ASIANA$ [0.20] | $KOREAN AIR$ [0.30] | $CAAC$ [0.25] | $Dragon Air$ [0.15]
- $Pronoun \rightarrow you$ [0.40] | I [0.60]

¹ 这是他在 1988 年 12 月 7 日在自然语言处理评测讨论会上的讲话。在 Palmer 和 Finin (1990) 描述这个讨论会时, 没有写下这段引文; 一些当时参加会议的人回忆, Jelinek 讲的话更为尖刻, 他说: “Every time I fire a linguist the performance of the recognizer improves.” (“每当我解雇一个语言学家, 语音识别系统的性能就会改善一些。”)

注意，所有从同一个非终极符号重写的规则的概率之和为 1。

如果分析的句子是有歧义的，PCFG 可给句子的每一个树一个概率。一个树 T 的概率应该等于从每一个结点 n 扩充的规则 r 的概率的乘积：

$$P(T) = \prod_{n \in T} p(r(n))$$

例如，句子 “Can you book ASIANA flights” 是有歧义的：一个意思是：“Can you book flights on behalf of ASIANA”；另一个意思是 “Can you book flights run by ASIANA”。它们的树分别如下：

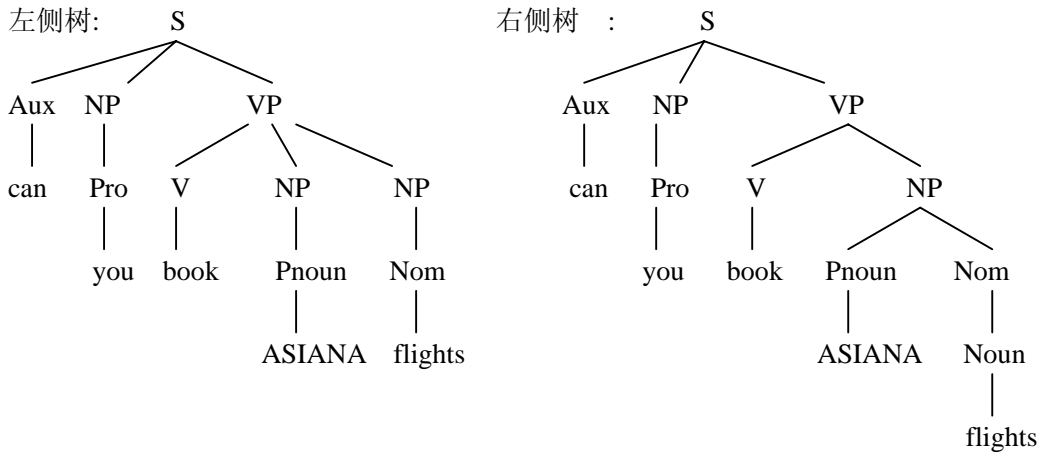


图 5 歧义句子“Can you book flights run by ASIANA”.的剖析树

左侧树的概率为：

- S → Aux NP VP [.15]
- NP → Pro [.40]
- VP → V NP NP [.05]
- NP → Nom [.05]
- NP → Pnoun [.35]
- Nom → Noun [.75]
- Aux → can [.40]
- NP → Pro [.40]
- Pro → you [.40]
- Verb → book [.30]
- Pnoun → ASIANA [.40]
- Noun → flights [.50]

右侧树的概率为：

- S → Aux NP VP [.15]
- NP → Pro [.40]
- VP → V NP [.40]
- NP → Nom [.05]
- Nom → Pnoun Nom [.05]
- Nom → Noun [.75]
- Aux → can [.40]
- NP → Pro [.40]
- Pro → you [.40]

Verb → book [0.30]

Pnoun → ASIANA [0.40]

Noun → flights [0.50]

左侧树的概率 $P(T_L)$ 为:

$$P(T_L) = .15 * .40 * .05 * .05 * .35 * .75 * .40 * .40 * .40 * .30 * .40 * .50 \\ = 1.5 \times 10^{-6}$$

右侧树的概率 $P(T_R)$ 为:

$$P(T_R) = .15 * .40 * .40 * .05 * .05 * .75 * .40 * .40 * .40 * .30 * .40 * .50 \\ = 1.7 \times 10^{-6}$$

可见, 右侧树的概率大于左侧树的概率, 所以, 右侧树可判定为正确的结果。

歧义消解算法从句子 S 的分析树 (我们把它们叫做 $\tau(S)$) 中选出最好的树 (我们把它叫做 T) 作为正确的分析结果。形式地说, 如果 $T \in \tau(S)$, 那么, 概率最大的数 $T(S)$ 将等于 $\operatorname{argmax} P(T)$ 。我们有:

$$T(S) = \operatorname{argmax} P(T)$$

■ 词汇化的概率上下文无关语法

PCFG 存在结构依存和词汇依存的问题

■ 结构依存问题:

CFG 语法假定, 规则左部的非终极符号进行重写时, 不依赖于其他的非终极符号。因此, 在 PCFG 中, 每一条规则都是独立的, 这样, 规则的概率才可以相乘。然而, 在英语中, 结点上规则的转写与结点在树形图中的位置有关。例如, 英语句子中的主语倾向于使用代词, 这是因为主语通常是表示主题或者旧信息, 而要援引旧信息时往往使用代词, 而不是代词的其他名词往往用于引入新信息。根据 Francis (1999) 的调查, 在 Switchboard 语料库中, 陈述句的主语有 31,021 个, 其中 91% 的主语是代词, 只有 9% 的主语是其他词。与此相反, 在 7,498 个宾语中, 只有 34% 是代词, 而 66% 是其他词。

主语: : **She** is able to take her baby to work with her.

: **My wife** worked until we had a family.

宾语: Some laws absolutely prohibit **it**.

All the people signed **applications**.

■ 词汇依存问题:

(1) PP 附着: 在英语句子中, 介词短语 PP 可以做中心动词短语 VP 的状语, 也可以做它前面名词短语 NP 的修饰语, 究竟是附着于 VP, 还是附着于 NP, 与词汇有关。例如, 在句子 “Washington sent more than 10,000 soldiers into Afghanistan”, PP “into Afghanistan” 或者附着于 NP (more than 10,000 soldiers), 或者是附着于动词(sent)。

在 PCFG 中。这种附着的判定要在下面的规则之间进行选择:

NP → NP PP (NP 附着)

和 VP → VP PP (VP 附着)

这两个规则的概率依赖于训练语料库:

语料库	NP 附着	VP 附着
AP Newswire (13 00 万词)	67%	33%
Wall Street Journal & IBM manuals	52%	48%

可以看出, NP 附着处于优先地位。

但是, 在我们上面的句子中, 介词短语 “into Afghanistan” 的正确附着应该是附着于动词

“sent”，这是因为动词 sent 往往要求一个表示方向的 PP，而介词短语 “into Afghanistan” 正好满足了这个要求。PCFG 显然不能处理这样的词汇依存问题。

(2) 并列结构的歧义：

句子 “dogs in houses and cats” 是有结构歧义的：

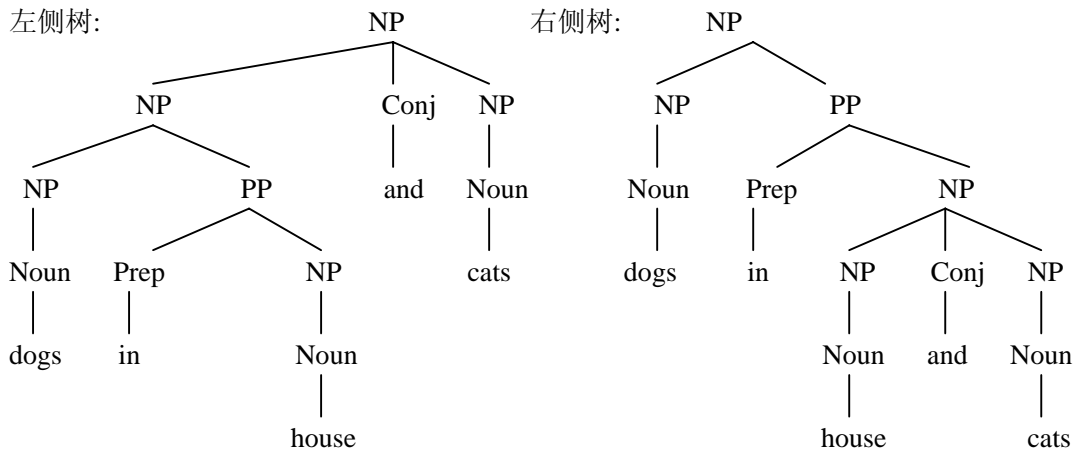


图 6 并列结构歧义

尽管在直觉上我们认为左侧树是正确的，但是，由于左右两侧的树所使用的规则是完全一样的，所以，它们的概率应该是一样的。

- NP → NP Conj NP
- NP → NP PP
- NP → Noun
- PP → Prep NP
- Noun → dogs | house | cats
- Prep → in
- Conj → and

在这种情况下，PCFG 将指派这两个树相同的概率，也就是说，PCFG 无法判定这个句子的歧义。为此，可以在 PCFG 中引入词汇信息来解决这些问题。

Charniak (1997) 提出了词汇中心语概率表示的方法。他的方法实际上是一种词汇语法 (lexical grammar.)，也叫做词汇化的概率上下文无关语法。在 Charniak 的概率表示中，剖析树的每一个结点要标上该结点的中心词 (head)。例如，句子 “Workers dumped sacks into a bin” 可表示如下：

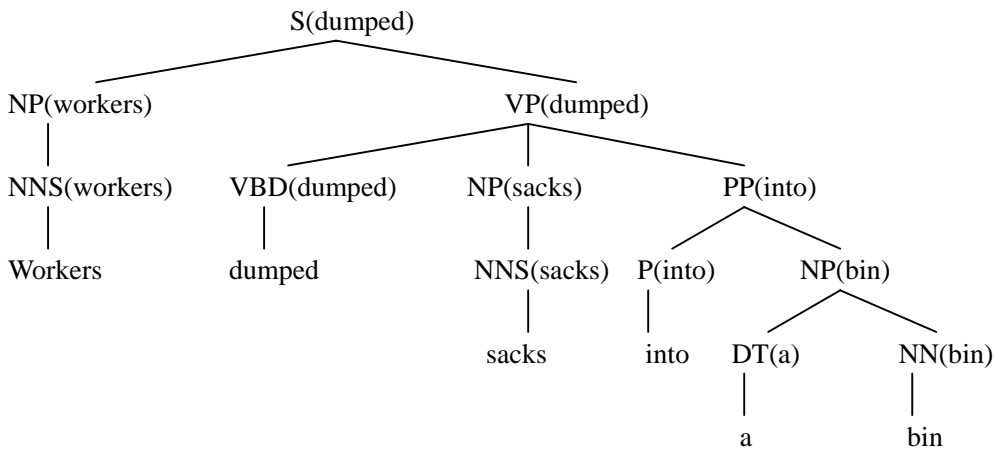


图 7 词汇化的剖析树

这时，词汇化的概率上下文无关语法的规则数目将比 PCFG 的规则多得多。例如，我们可以

有如下的规则，规则中既包括概率，也包括词汇信息：

VP(dumped) → VBD(dumped) NP(sacks) PP(into) [3X10⁻¹⁰]

VP(dumped) → VBD(dumped) NP(cats) PP(into) [8X10⁻¹¹]

VP(dumped) → VBD(dumped) NP(hats) PP(into) [4X10⁻¹⁰]

VP(dumped) → VBD(dumped) NP(sacks) PP(above) [1X10⁻¹²]

句子也可以被剖析为另一个树形图：

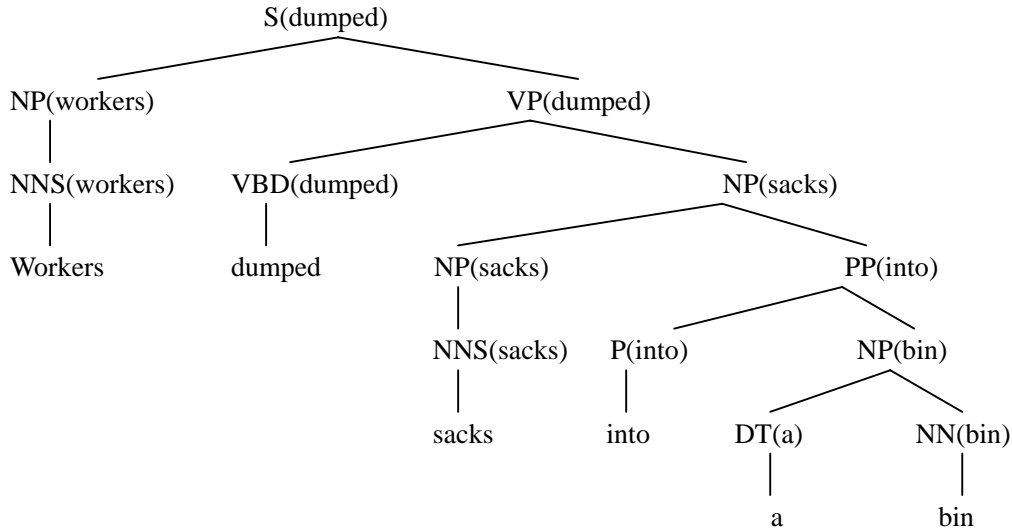


图 8 不正确的剖析树

如果 VP(dumped) 重写为 VBD NP PP，那么，我们可以得到正确的剖析树。如果我们把 VP(dumped) 重写为 VBD NP，那么，就得到上面的这个不正确的剖析树。

我们可以根据 Penn Tree-bank 中的 Brown 语料库来计算这种词汇化规则的概率：

第一个规则的概率为：

$$P(\text{VP} \rightarrow \text{VBD NP PP} | \text{VP, dumped}) = \frac{C(\text{VP(dumped)} \rightarrow \text{VBD NP PP})}{\sum_{\beta} C(\text{VP(dumped)} \rightarrow \beta)} = \frac{6}{9} = .67$$

第二个规则从不在 Brown 语料库中出现，因为“dump”这个动词要求指明动作所到达的新位置，因此，如果它后面没有介词短语，就是不合理的。

$$P(\text{VP} \rightarrow \text{VBD NP} | \text{VP, dumped}) = \frac{C(\text{VP(dumped)} \rightarrow \text{VBD NP})}{\sum_{\beta} C(\text{VP(dumped)} \rightarrow \beta)} = \frac{0}{9} = 0.$$

在实际的应用中，如果概率出现零值，一般都要进行平滑（smoothing），这里我们不考虑平滑问题。

我们也可以用同样的方法来计算中心词的概率。

在正确的剖析树中，结点 PP 的母亲结点(X)是中心词“dumped”，在不正确的剖析树中，结点 PP 的母亲结点(X)是中心词“sacks”

根据 Penn Tree-bank 的 Brown 语料库，我们有

$$P(\text{into} | \text{PP, dumped}) = \frac{C(\text{X(dumped)} \rightarrow \dots \text{PP (into)} \dots)}{\sum_{\beta} C(\text{X(dumped)} \rightarrow \dots \text{PP} \dots)} = \frac{2}{9} = .22$$

$$P(\text{into} | \text{PP}, \text{sacks}) = \frac{C(X(\text{sacks}) \rightarrow \dots \text{PP}(\text{into}) \dots)}{\sum_{\beta} C(X(\text{sacks}) \rightarrow \dots \text{PP} \dots)}$$

$$= 0/0 = ?$$

可见，通过计算 PP 结点的母亲结点的概率，也可以判断 into 修饰 dumped 的概率比修饰 sacks 的概率大。

PCFG 和词汇化 PCFG 对于规则方法和统计方法的结合，进行了有成效的探索。

机器翻译系统中也开始采用规则和统计想结合的开发策略。

王野翊发现IBM系统中，词对齐的错误严重影响翻译质量。他在CMU研制的英德口语机器翻译系统，提出基于结构的翻译模型。这个模型包括一个粗对齐模型(rough alignment)和一个细对齐模型(detailed alignment)。源语言和目标语言的短语首先通过粗对齐模型进行对齐，然后对于短语内部的单词再通过细对齐模型进行对齐

在训练语料库上，通过基于互信息的双语词语聚类 and 短语归并反复迭代，得到一组基于词语类聚的短语规则，再利用这组短语规则对于句子的短语进行分析。结构知识的融入降低了系统11%的错误率，缓解了由于口语数据的缺乏而导致的数据稀疏问题。由此可见，在基于规则的方法中引入统计方法，是有积极意义的。

香港科技大学 Wu Dekai 提出随机可逆转换语法 (Stochastic Inversion Transformation Grammar, 简称 SITG)，他主张在原语分析的同时，也用同一套机制来生成目标语言，也就是说，原语的分析 and 目标语言的生成共享相同的机制，这实际上是一种平行概率语法。处理过程中使用带有概率的可逆转换语法规则，类似于 PCFG 的规则。所以，这是一种把规则和统计相结合的方法。这种方法对于语料库的加工和机器翻译都是很有意义的。

1.6 语音机器翻译系统

■ AT&T 公司的语音机器翻译系统

AT&T 公司的 Alshawaki (1998) 等开发的语音翻译系统由语音识别、机器翻译、语音合成三部分组成。他们在机器翻译部分采用的算法非常独特，这实际上是一个基于平行概率语法的机器翻译系。

该系统的总体工作流程如下图所示：

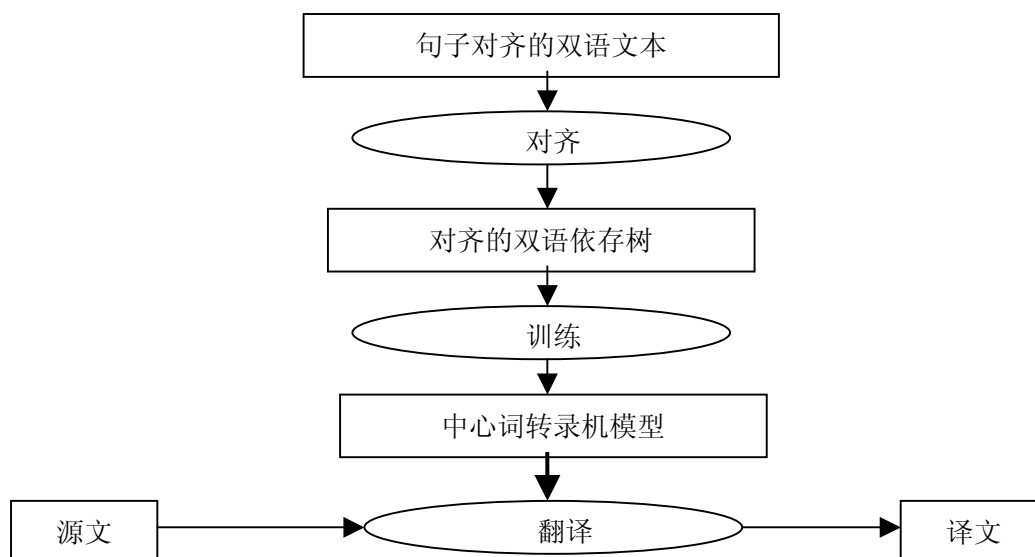


图 9 AT&T 的语音机器翻译系统

所采用的核心的数学工具是“中心词转录机”(Head Transducer, 简称 HT)。在这个系统中,系统的所有翻译知识(包括词典、分析规则和生成规则)都用一组带概率的 HT 来表示。这些知识完全从语料库中自动获取,获取的过程无需人工干预,但获取的结果(就是 HT)很直观,可以由人进行调整。HT 的表示是完全基于词的,不采用任何词法、句法或语义标记。

整个知识获取的过程实际上就是一个双语语料库结构对齐的过程。句子的结构用依存树表示(但依存关系不作任何标记)。他们经过一番公式推导,把一个完整的双语语料库的分析树构造和对齐的过程转化成了一个数学问题的求解过程。这个过程可用一个算法高效地实现。得到对齐的依存树后,很容易就训练出一组带概率的 HT,这样也就等于得到了一个机器翻译系统。

这种方法的主要特点是:1.训练可以全自动进行,效率很高,由一个双语句子对齐的语料库可以很快训练出一个机器翻译系统;2.不使用任何人为定义的语言学标记(如词性、短语类、语义类等等),无需任何语言学知识;3.训练得到的参数包含了句子的深层结构信息,这一点比 IBM 的统计语言模型更好。

这种方法比较适合于语音翻译这种领域比较受限,词汇集较小的场合,对于大规模的文本翻译并不合适。但这种做法对我们开拓思路还是非常有借鉴意义的。

■ Verbmobil 系统

Verbmobil是由德国教育与研究部(BMBF)资助的一个为期8年(1993-2000)语音机器翻译项目,涉及三种语言(德语、英语、日语)的双向翻译。世界三大洲的31个研究机构、369名科学家和919名学生(硕士生、博士生和博士后)参与了这个项目的研究。Verbmobil的目的在于“通过工业及科学界尽可能多的分支领域的合作与集中,在下一个世纪的语言技术及其经济应用领域中为德国谋取国际领先地位”。Verbmobil制定了1993-2001年的研制计划,其中自1993年至1996年的第一阶段计划吸收了德国、美国和日本的32个企业和高等学校的成员参加,政府投入资金4690万马克,企业投入资金310万马克,第一阶段的目标是建立非特定人的、面向会面安排交谈的口语语音翻译系统。

Verbmobil 系统与我们所熟悉的文本翻译系统不同之处主要体现在:

--语音处理:要进行语音识别和语音合成。该系统的目标很高,实现了 GSM 语音条件下的自动翻译,除了一开始拨打 Verbmobil 语音服务电话以外,整个系统的服务可完全用 GSM 电话通过语音方式实现,无需任何按键操作;系统具有语音自适应能力,一开始使用与说话者无关的语音识别模块,通过一段时间对话后,自动适应说话者的口音,提高识别正确率;

--处理自然的语音:要考虑现实口语中的各种复杂现象,如停顿、重复、修正、漏词等等;要建立对话模型,理解句子的语义,并考虑上下文进行翻译,甚至要猜测说话者的意图;

系统开发面向三个商务领域:约会安排、旅行计划和远程 PC 维护,复杂程度依次增加,如下表所示:

场景一: 约会安排	场景二: 旅游计划	场景三: 远程 PC 维护
何时?	何时? 何地? 如何?	何物? 何时? 何地? 如何?
关注时间性表达	关注时间性和空间性表达	一个词汇受限的子语言
词汇量: 2500/6000	词汇量: 7000/10000	词汇量: 15000/30000

可以看出,虽然领域受到限制,但这还是一个相当复杂的系统,处理的问题面极其广泛。从以下的系统结构图中可以看到这点:

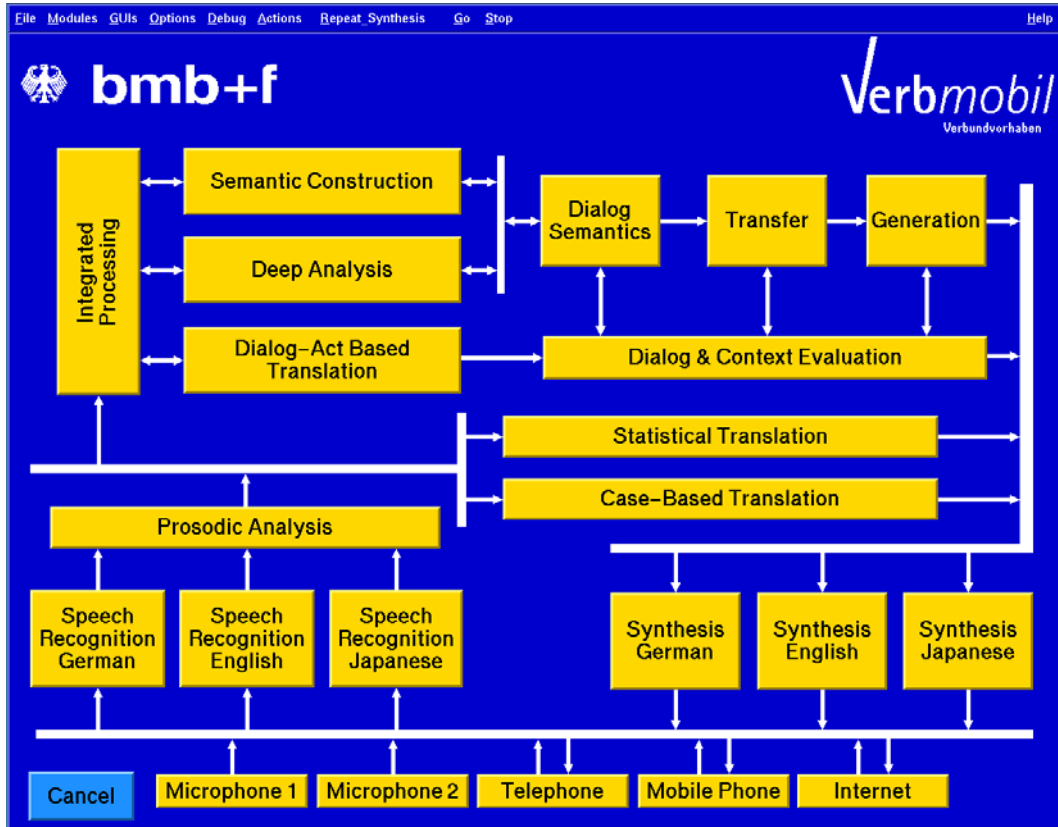


图 10 Verbmobil 系统

可以看出，语音处理领域和自然语言处理领域中几乎所有的各种技术都在这个系统中有所反映。整个系统由 69 个互相交互的模块构成。其中用到的自然语言处理技术包括：组块分析、概率 LR 分析、HPSG 分析、对话行为（Dialog Act）分析、基于统计的翻译、基于子串的翻译、基于模板的翻译、基于模板的转换、语义分析、上下文相关歧义的消解、基于规划的话语生成，等等。

系统采用一种多黑板结构进行模块之间的数据交互，模块之间不能直接通信。黑板结构还有利于各个模块之间的并行执行。总共采用了 198 个黑板结构用于 69 个不同模块之间的通讯。一种叫做 VIT（Verbmobil Interface Terms）的数据结构在中心黑板的深层处理中用作深层的语义表示形式。

句法分析阶段采用了多引擎技术，利用了三个分析引擎，分别是：组块分析器（Chunk Parser）、概率 LR 分析器和 HPSG 分析器。其中组块分析器鲁棒性最好，但结果质量最差，HPSG 分析器鲁棒性最差，但结果质量最好。

系统管理上采用了严格的软件工程方法，保证了项目的顺利实施。

组织上，Verbmobil 强调各个研究组之间的竞争，通过频繁的目标评价。项目中不同的团队对于每个特定的目标都提出了竞争的解决方案，通过正式的评价，从中选择最好的方案或者与次好的方案合并以改进总体的效果。

多种基准的测试以及大规模端对端评价实验令人信服地表明，Verbmobil 的最终版本系统中达到了所有的预定目标，有些目标甚至被超越。在大规模翻译实验中，正确翻译率达到大约 80%，在真实用户的端对端测试中，90%的对话任务获得成功。

1.7 机器翻译的存在是不可否认的现实

机器翻译的研制已经有 50 多年的历史了，由于自然语言的复杂性，机器翻译的译文难以做到完满的程度，但是，机器翻译的存在是一个不可否认的现实，你喜欢它也好，你厌恶它也好，它总是存在的。它决不会因为你的喜好而马上达到全自动高质量，也不会因为你的厌恶而放弃自身的存在以至于自动推出历史舞台。这里我们列举几个事实：

- 美国 Ohio 州 Dayton 市的联邦翻译部多年来一直没有间断过机器翻译工作，他们为美国科学工作者提供俄英科技文献的机器翻译服务，每天翻译数百个单词的文章，翻译数量虽然不大，但是多年持之以恒。
- 日本电子工业振兴学会评估日本翻译市场的规模是每年 80 亿美元，翻译市场如此之大，对机器翻译的要求非常迫切。

在这种情况下，我们有必要重新定义“成功的机器翻译”的概念。定义时应该考虑到如下的因素：

- 对机器翻译的期望值不要太高，应该接受质量低的译文，容忍低质量的译文。
- 不要求机器翻译一定要全自动。
- 机器翻译要量力而行，不要勉强翻译那些机器根本无法处理的过分复杂的文本。
- 机器翻译是有用的，但机器翻译决不是万能的。

如果我们采取实事求是的态度来看待“成功的机器翻译”，我们就不会对机器翻译提出过于苛刻的要求。

3. 对于机器翻译研究的一些建议

3.1 语义分析问题

机器翻译中是应该研究语义的，但是，语义处理如此之困难，在某些情况下，我们能不能避开语义处理这个难题呢？有很多学者主张。对于语义处理，如果能够避开就尽量避开，不要正面对抗，而要采用灵活的策略来处理它。下面是他们的一些真知灼见，我只录其英文，请读者玩味：：

- PP 附着问题，有人提出：“Why waste time detecting and representing the meaning of the input string when the target language correlate is always the same?” 这意味着，如果两种语言都存在 PP 附着问题，我们可以采用以“歧义对歧义”的翻译策略，不一定要在原语的歧义分析时打硬战，避开这场硬战可能还会得到更好的译文。
- Ben Ari 等提出：“It must kept in mind that the translation process does not necessarily require full understanding of the text. Many ambiguities may be preserved during translation, and thus should be presented to the user (human translator) for resolution.”他们不主张在充分理解文本之后才进行机器翻译，歧义问题可以由人来判定。
- Isabelle and Bourbeau 提出：“Sometimes, it is possible to ignore certain ambiguities, in the hope that the same ambiguities will carry over in translation. This is particularly true in system like TAUM-aviation that deals with only one pair of closely related languages within a severely restricted sub-domain. The difficult problem of prepositional phrase attachment, for example, is frequently bypassed in this way. Generally speaking, however, analysis is aimed at producing an unambiguous intermediate representation.”根据 TAUM 系统的经验，他们主张忽略文本中的某些歧义。

3.2 改变机器翻译的接受阈值

机器翻译的文本一般都不用于出版，因此，机器翻译的译文质量不应该用出版物的质量标准来衡量，只要能够传达原文的内容，机器翻译的译文就应该是可以接受的。因此，我们应该明确地降低机器翻译的译文质量的接受阈值，不能对机器翻译提出过于苛刻的要求。

法国机器翻译专家 Ch. Boitet 认为，从使用的目的来看，机器翻译可以分为 4 种，

- 用于浏览的机器翻译 (For the watcher, 简称 MT-W): 用户愿意接受粗糙的译文, 也不愿意一无所获。;
- 用于修改的机器翻译 (For the reviser, 简称 MT-R): 机器翻译的目的在于给用户 提供“草稿”, 以使用户修改。
- 用于翻译的机器翻译 (For the translator, 简称 MT-T): 机器翻译系统为翻译人员 提供在线词典、参考译文的实例, 帮助翻译人员进行翻译。
- 用于作者的机器翻译 (For the author, 简称 MT-A): 作者在翻译时, 由机器翻译 提供译文中可能产生的歧义, 以便得到较好的译文。

这 4 种类型实际上也反映了机器翻译发展的不同阶段, 随着机器翻译技术的发展, 机器翻译的接受阈值逐渐提高, 从而得到越来越好的译文。这是一种比较实际的想法。

3.3 提倡半自动的机器翻译

在机器翻译的目前水平下, 采用人机结合的方式来进行机器翻译是可行的。目前的主要方式有:

- 采用译后编辑对译文进行修改: 如 SYSTRAN 系统;
- 建立人机交互的计算机环境, 开发翻译工作站 (translator's workstation);
- 建立翻译记忆库 (Translation memory): 存储已经翻译成功的语言资料, 尔后的 翻译主要是查询这个翻译记忆库, 如 Trados 系统。

3.4 限制原语的歧义

- 选择一个范围比较窄的领域, 有针对性地开发机器翻译系统有可能限制原语的歧 义。如加拿大的 Montreal 大学开发的 TAUM-METEO 系统, 由于专业领域选择 得当, 所用的词汇大约只有 1500 个不同的单词, 而且半数是地名, 系统的词汇 歧义 (多义词) 很小, 尽管是多义词, 但是由于领域的限制, 在特定的领域中也 没有歧义。当然, 要选择到像 TAUM-METEO 这样的领域需要精心地调查, 也不 是一件容易的事情。我国机器翻译怎样去发现一个像加拿大 TAUM-METEO 这样 的领域, 是需要我们认真研究的。
- 译前编辑有针对性地编辑那些可能引起歧义的原文片段, 尽量减少机器翻译过程中 的歧义。

3.5 加强机器翻译的评测工作, 通过评测带动学科的发展

- 约翰霍普金斯大学 (JHU) 于 1999 年夏季召开了一个研讨班, 重复了 IBM 的统计机器翻 译工作, 开发了一个源代码公开的统计机器翻译软件工具包 Egypt, 这个工具包公开散 发, 并使用这个工具包作为试验性的平台进行了一系列的试验。在这个研讨班上, 构造 了一个捷克语—英语的机器翻译系统。他们进行基准评测: 包括客观评测 (评测统计模 型的困惑度) 和主观评测 (翻译质量的人工判断)。在研讨班的最后一天, 在一天之内 构造出一个新语言对的机器翻译系统。这个统计机器翻译软件 Egypt 工具包有如下 的若干个模块
 - Giza: 用于在双语语料库中进行参数训练, 获取统计知识
 - Decoder 解码器: 用于具体的翻译过程, 进行解码 (翻译)

- Cairo: 翻译系统的可视化界面, 可以观察双语对齐的过程和机器翻译的解码过程
- Whittle: 语料库预处理的工具

Egypt是免费的工具包, 其源代码可以在网上自由下载。

这样通过评测来推动机器翻译是有效果的。我国在 863 计划中, 也进行过 3 次评测, 但是后来中断了, 我希望这样的评测活动能够坚持下去, 通过评测来推动机器翻译事业的发展。

3.6 把机器翻译与 NLP 的其他领域的应用结合起来

信息检索系统的检索文本不一定要求高质量的文本, 如果机器翻译文本包含的信息是正确的, 尽管译文粗糙, 也可以作为信息检索的对象。信息检索并不要求对于文本中的每一个单词都进行精确的分析。因此, 如果能够把机器翻译和双语言信息检索结合起来, 有可能推广机器翻译的应用领域。

4. 机器翻译还是一个不成熟的研究领域

目前, 在机器翻译领域中, 存在着各种各样的机器翻译, 其中有:

- 全自动机器翻译/机器辅助翻译 (译后编辑)
- 高质量的译文/粗糙译文
- 双语言机器翻译/多语言机器翻译
- 不限制输入的原文/控制输入的原文 (译前编辑)
- 文本翻译/语音翻译/文本-语音翻译
- 全文机器翻译/文摘机器翻译/题录机器翻译
- 基于规则的机器翻译/基于语料库的机器翻译
- 中间语言的机器翻译/转换的机器翻译/词对词机器翻译
- 通用机器翻译/面向特定领域的机器翻译
- 基于统计机器翻译/基于实例的机器翻译

可以说, 机器翻译的研究出现了百花齐放的局面。

但是, 机器翻译还是一个不成熟的研究领域, 原因如下:

- 机器翻译对于任务的依赖性很强, 如果想得到好的翻译结构, 必须慎重地选择翻译的文本。
- 要达到全自动高质量的机器翻译, 还存在着严重的知识空缺的瓶颈问题
- 机器翻译的文本分析, 目前只是针对一个单一的句子来进行, 还没有考虑句子与句子之间的关系, 计算语用学的研究非常单薄。
- 机器翻译系统的鲁棒性还很差, 如果输入文本中存在单词遗漏和错误拼写的单词, 如果单词和短语的语义超出词典所包含的范围, 都会导致系统的失败。
- 机器翻译系统过分依赖于所翻译的语言对。如果语言对比较接近, 翻译效果就好一些, 如果语言对差别很大, 翻译效果就很差。
- 机器翻译的基础理论研究还很差。
- 在机器翻译的研究中, 劳动密集型的研究较多, 资源密集型的研究较少。

我们应当继续努力, 把机器翻译的研究推向一个新的阶段。

参 考 文 献

1. Ben Ari et al, Translation ambiguity rephrased, In Proceedings of the 2nd International Conference on Theoretical and methodological Issues in Machine Translation of Natural Languages, 1988.
2. Bonnie J. Dorr, Pamela W. Jordan, John W. Benoit, A survey of Current Paradigms in Machine

Translation, <Advances in Computers>, Volume 49, 1999, Academic Press.

3. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., A maximum Entropy Approach to Natural Language Processing, Computational Linguistics, Volume 22, No.1
4. Charniak, E., Statistical Parsing with a Context-free Grammar and Word Statistics. In Fourteenth national Conference on Artificial Intelligence Providence, Rhode Island.
5. Isabelle and Bourbeau, TAUM-AVIATION: its technical features and some experimental results, Computational Linguistics, 11(1),1985
6. Nirenburg, S., Yorick Wilks, Machine Translation, <Advances in computers>, Volume 52, 2000, Academic Press.
7. 刘群, 俞士汶, 机器翻译技术综述及面向新闻领域的汉英机器翻译系统, 北京大学计算语言学研究所资料, 2002 年。
8. 冯志伟, 计算语言学探索, 黑龙江教育出版社, 2002 年。
9. 冯志伟, 计算语言学基础, 商务印书馆, 2001 年。