计算语言学

第8讲 机器翻译I

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

大纲

- 机器翻译的历史
- 机器翻译的分类
- 机器翻译的范式
- 机器翻译的基本策略
- 基于规则的机器翻译方法
- 基于实例的机器翻译方法

（本节讲义很多内容直接取自冯志伟教授的一份讲义，特此表示感谢）

机器翻译的历史

- W. J. Hutchens, latest Development in MT Technology: Beginning a New Era in MT Research. In : Proceedings of Machine Translation Summit-IV, Kobe, Japan, 1993.
- 冯志伟，自动翻译，上海知识出版社，1987年。
- 冯志伟，自然语言机器翻译新论，语文出版社，1994年。
- 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996年。

机器翻译的萌芽期（1）

- 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。

机器翻译的萌芽期（2）

- 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。

机器翻译的萌芽期（3）

- 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。

机器翻译的草创期（1）

- 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。”

机器翻译的草创期（2）

- 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，他除了提出各种语言都有许多共同的特征这一论点之外，还有两点值得我们注意：
  第一，他认为翻译类似于解读密码的过程。他说：“当我阅读一篇用俄语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读时，我是在进行解码。”
  第二，他认为原文与译文“说的是同样的事情”，因此，当把语言A翻译为语言B时，就意味着，从语言A出发，经过某一“通用语言”（Universal Language）或“中间语言”（Interlingua），然后转换为语言B，这种“通用语言”或“中间语言”，可以假定是全人类共同的。
- 由于学者的热心倡导，实业界的大力支持，美国的机器翻译研究一时兴盛起来。1954年，美国乔治敦大学在国际商用机器公司（IBM公司）的协同下，用IBM-701计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

机器翻译的萧条期

- 1964年，美国科学院成立语言自动处理谘询委员会（Automatic Language Processing Advisory Committee，简称ALPAC委员会），调查机器翻译的研究情况，并于1966年11月公布了一个题为《语言与机器》的报告，简称ALPAC报告，对机器翻译采取否定的态度，报告宣称：“在目前给机器翻译以大力支持还没有多少理由”； 报告还指出，机器翻译研究遇到了难以克服的“语义障碍”（semantic barrier）。
- 在ALPAC报告的影响下，许多国家的机器翻译研究低潮，许多已经建立起来的机器翻译研究单位遇到了行政上和经费上的困难，在世界范围内，机器翻译的热潮突然消失了，出现了空前萧条的局面。

机器翻译的复苏期（1）

- 尽管在萧条时期，法国、日本机器翻译研究的历史和现状加拿大等过，仍然坚持着机器翻译研究，于是，在七十年代初期，机器翻译又出现了复苏的局面。
- 在这个复苏期，研究者们普遍认识到，原语和译语两种语言的差异，不仅只表现在词汇的不同上，而且，还表现在句法结构的不同上，为了得到可读性强的译文，必须在自动句法分析上多下功夫。

机器翻译的复苏期（2）

- 早在1957年，美国学者英格维（V. Yingve）在《句法翻译的框架》（Framework for syntactic translation）一文中就指出，一个好的机器翻译系统，应该分别地对原语和译语都作出恰如其分的描写，这样的描写应该互不影响，相对独立。英格维主张，机器翻译可以分为三个阶段来进行。
  - 第一阶段：用代码化的结构标志来表示原语文句的结构；
  - 第二阶段：把原语的结构标志转换为译语的结构标志；
  - 第三阶段：构成译语的输出文句。

机器翻译的复苏期（3）

- 这个时期机器翻译的另一个特点是语法（grammar）与算法（algorithm）分开。
- 早在1957年，英格维就提出了把语法与“机制”（mechanism）分开的思想。英格维所说的“机制”，实质上就是算法。所谓语法与算法分开，就是要把语言分析和程序设计分开，程序设计工作者提出规则描述的方法，而语言学工作者使用这种方法来描述语言的规则。语法和算法分开，是机器翻译技术的一大进步，它非常有利于程序设计工作者与语言工作者的分工合作。

机器翻译的复苏期（4）

- 这个复苏期的机器翻译系统的典型代表是法国格勒诺布尔理科医科大学应用数学研究所（IMAG）自动翻译中心（CETA）的机器翻译系统。这个自动翻译中心的主任沃古瓦（B. Vauquois）教授明确地提出，一个完整的机器翻译过程可以分为如下六个步骤：
  （1）原语词法分析
  （2）原语句法分析
  （3）原语译语词汇转换
  （4）原语译语结构转换
  （5）译语句法生成
  （6）译语词法生成

  其中，第一、第二步只与原语有关，第五、第六步只与译语有关，只有第三、第四步牵涉到原语和译语二者。
- 这就是机器翻译中的“独立分析-独立生成-相关转换”的方法。他们用这种研制的俄法机器翻译系统，已经接近实用水平。

机器翻译的复苏期（5）

- 他们还根据语法与算法分开的思想，设计了一套机器翻译软件ARIANE-78，这个软件分为ATEF，ROBRA，TRANSF和SYGMOR四个部分。语言工作者可以利用这个软件来描述自然语言的各种规则。
- ATEF是一个非确定性的有限状态转换器，用于原语词法分析，它的程序接收原语文句作为输入，并提供出该文句中每个词的形态解释作为输出；
- ROBRA是一个树形图转换器，它的程序接收词法分析的结果作为输入，借助语法规则对此进行运算，输出能表示文句结构的树形图；ROBRA还可以按同样的方式实现结构转换和句法生成；
- TRANSF可借助与双语词典实现词汇转换；
- SYGMOR是一个确定性的树-链转换器，它接收译语句法生成的结果作为输入，并以字符链的形式提供出译文。

机器翻译的复苏期（6）

- 美国斯坦福大学威尔克斯（Y.A. Wilks）提出了“优选语义学”（preference semantics），并在此基础上设计了英法机器翻译系统。
- 这个系统特别强调在原语和译语生成阶段，都要把语义问题放在第一位，英语的输入文句首先被转换成某种一般化的通用的语义表示，然后再由这种语义表示生成法语译文输出。
- 由于这个系统的语义表示方法比较细致，能够解决仅用句法分析方法难于解决的歧义、代词所指等困难问题，译文质量较高。

机器翻译的繁荣期

- 本世纪七十年代末，机器翻译进入了它的第三个时期--繁荣期（1976年--现在）。
- 繁荣期的最重要的特点，是机器翻译研究走向了实用化，出现了一大批实用化的机器翻译系统，机器翻译产品开始进入市场，变成了商品，由机器翻译系统的实用化引起了机器翻译系统的商品化。

机器翻译的分类 1

- 理想的机器翻译
  - 全自动高质量，FAHQ MT
    Full Automatic High Quality Machine Translation
- 按人机关系分类
  - 全自动机器翻译，FAMT
    Full Automatic Machine Translation
  - 人助机译，HAMT
    Human Assisted Machine Translation
  - 机助人译，CAT
    Compute-Aided Translation

机器翻译的分类 2

- 按应用方式分类
  - 信息分发型 MT for dissemination
    - 要求高质量，不要求实时
    - 采用人机互助，或者受限领域、受限语言等方式提高翻译质量
  - 信息吸收型 MT for assimilation
    - 不要求高质量，要求方便、实时
    - 翻译浏览器、便携式翻译设备、……

机器翻译的分类 3

- 按应用方式分类（续）
  - 信息交流型 MT for interchange
    - 不要求高质量，通常要求实时，语言随意性较大
    - 语音翻译、网络聊天翻译、电子邮件翻译
  - 信息存取型 MT for information access
    - 将机器翻译嵌入到其他应用系统中
    - 跨语言检索、跨语言信息抽取、跨语言文摘、跨语言非文本数据库的检索……

口语机器翻译系统（1）

- ATR-ITL口语翻译系统：近年来，国外开始自动翻译电话的研究，在日本关西地区成立了自动电话研究所（Advanced Telecommunications Research Institute International – Interpreting Telecommunications Research laboratories，简称 ATR-ITL）,其目的在于把语音识别、语音合成技术用于机器翻译中，实现语音机器翻译。1989年，日本ATR研制了SL-TRANS系统。
- SpeechTrans系统和JANUS系统：由美国卡内基-梅隆大学(CMU)研制。
- KITANO系统：90年代初期，日本学者北野(Kitano)在京都大学期间，使用大规模并行计算，采用基于实例的方法进行语音翻译实验，证明了毫秒级的实时口语语音翻译是可实现的。

口语机器翻译系统（2）

- Verbmobil计划：由德国联邦政府教育、科学、研究与技术部(BMBF)支持，其目的在于“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。

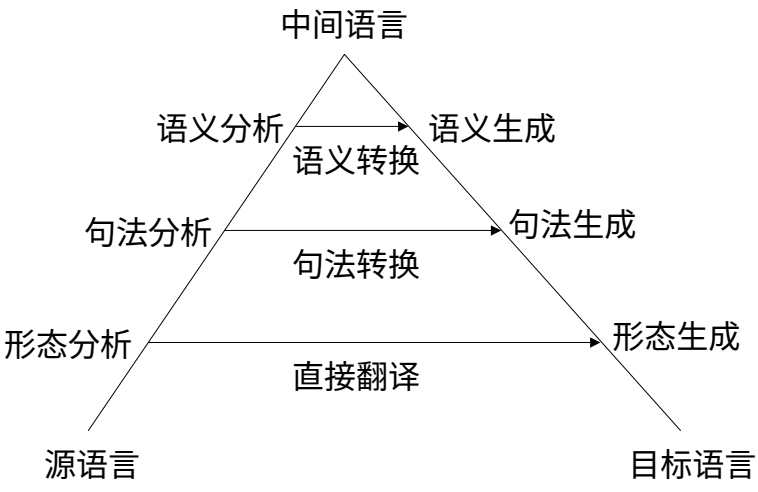
口语机器翻译系统（3）

- C-STAR计划：1991年成立了国际口语翻译联盟(Consortium for Speech Translation Advanced Research，简称C-STAR)。 C-STAR是一个以口语语音翻译为基本研究目标的国际合作组织，由来自12个国家的20个成员组成。
- 核心成员有来自7个国家7个单位：美国的卡内基-梅隆大学(CMU)、日本的ATR-ITL、德国的卡尔斯鲁尔大学UKA (University Karlsruhe)、法国格勒诺布尔大学自动翻译研究中心GETA-CLIPS、意大利的科学技术研究所ITC-IRST、韩国的高级网络服务技术部ETRI、中国科学院自动化研究所国家模式识别重点实验室(NLPR)。其他成员有德国西门子公司(Siemens)、香港科技大学等。
- C-STAR把多种语言的口语直接翻译作为一个科学工程来进行，通过建立平台和演示来推动口语语音翻译技术的迅速发展，使C-STAR成为国际口语翻译技术转向工业应用的摇篮，以扫除人类的语言障碍。
- 作为C-STAR核心成员的中国科学院自动化所NLPR已经建立了口语翻译的试验系统的相关平台，完成了一个面向会面安排的汉英口语语音机器翻译原型系统EasySchedule，正在开发可初步实用的汉英口语语音机器翻译系统。

互连网络与机器翻译

- 词典容量大而不失其准：由于网络上文本涉及面广，词汇十分丰富，网络翻译系统的词典容量都很大，至少可以帮助人们查询不认识的生词，弄清生词的准确含义。
- 翻译速度快而不失其要：便于在网上快速浏览并查找所需要的信息，了解网上信息的梗概要略，译文具有可读性。
- 译文质量粗而不失其信：译文能传达原文的意思，以“信”为首先的追求目标，而不要求做到译文的“达”和“雅”。
- 翻译方式多而不失其巧：既可以使用Web浏览器将原文下载到PC机上进行翻译，也可以在网络上直接控制进行翻译，也可以使用proxy代理服务器代表客户机传送服务请求，通过翻译软件在Web浏览器(Navigator 1.0或2.01，Internet Explorer 2.0)上把源语言直接翻译为目标语言，还可以仅只查词典，翻译方式多样而巧妙，以应不同用户的要求。
- 文本格式严而不失其便：译文尽量保持原文的“超文本”特点，满足HTML超文本置标语言的要求，便于用户在网络中畅游。

机器翻译的范式（Paradigm）



直接翻译方法

- 通过词语翻译、插入、删除和局部的词序调整来实现翻译，不进行深层次的句法和语义的分析，但可以采用一些统计方法对词语和词类序列进行分析
- 早期机器翻译系统常用的方法，近期IBM提出的统计机器翻译模型也可以认为是采用了这一范式
- 著名的机器翻译系统Systran早期也是采用这种方法，后来逐步引入了一些句法和语义分析

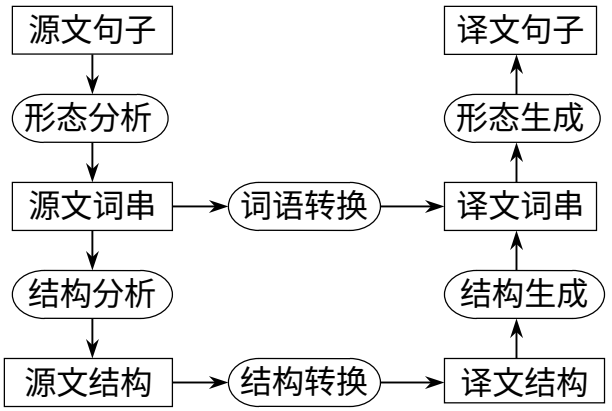
转换方法 1

- 整个翻译过程分为“分析”、“转换”、“生成”三个阶段；
- 分析：源语言句子➔源语言深层结构
  - 相关分析：分析时考虑目标语言的特点
  - 独立分析：分析过程与目标语言无关
- 转换：源语言深层结构➔目标语言深层结构
- 生成：目标语言深层结构➔目标语言句子
  - 相关生成：生成时考虑源语言的特点
  - 独立生成：生成过程与源语言无关

转换方法 2

- 理想的转换方法应该做到独立分析和独立生成，这样在进行多语言机器翻译的时候可以大大减少分析和生成的工作量；
- 转换方法根据深层结构所处的层面可分为：
  - 句法层转换：深层结构主要是句法信息
  - 语义层转换：深层结构主要是语义信息
- 分析深度的权衡
  - 分析的层次越深，歧义排除就越充分
  - 分析的层次越深，错误率也越高

转换方法 3



基于转换方法的翻译流程

句法层面的转换方法 1

她把一束花放在桌上。⇨ She put a bunch of flowers on the table.

⇩

切分／标注

⇩

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

⇩

标注排歧

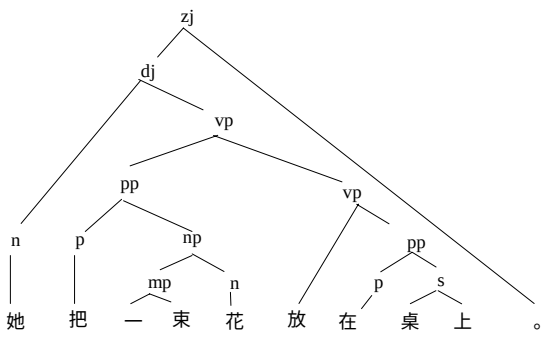
⇩

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

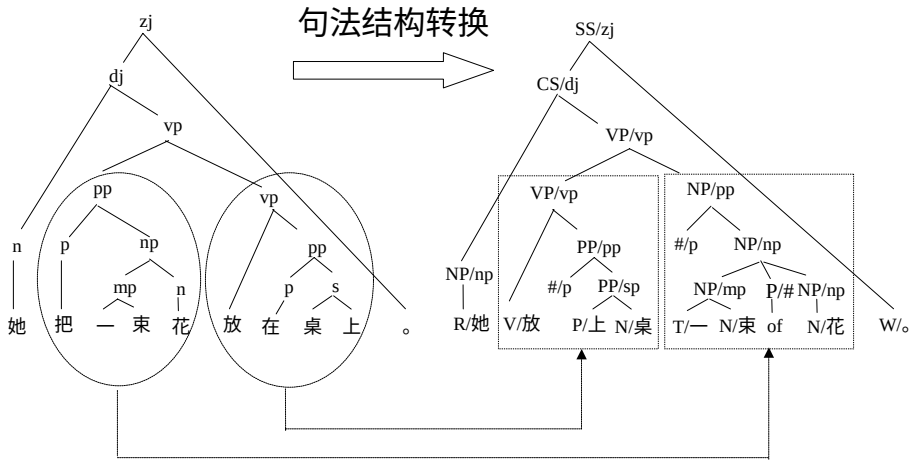
句法层面的转换方法 2

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

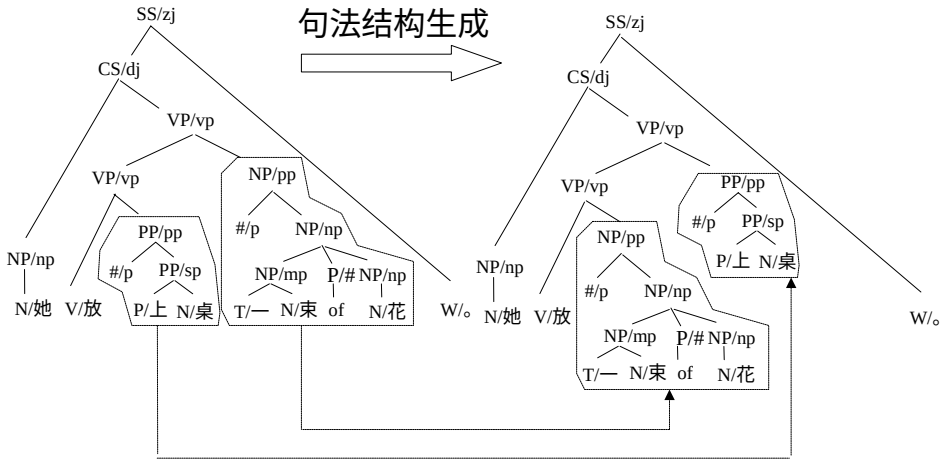
⇩ 句法分析



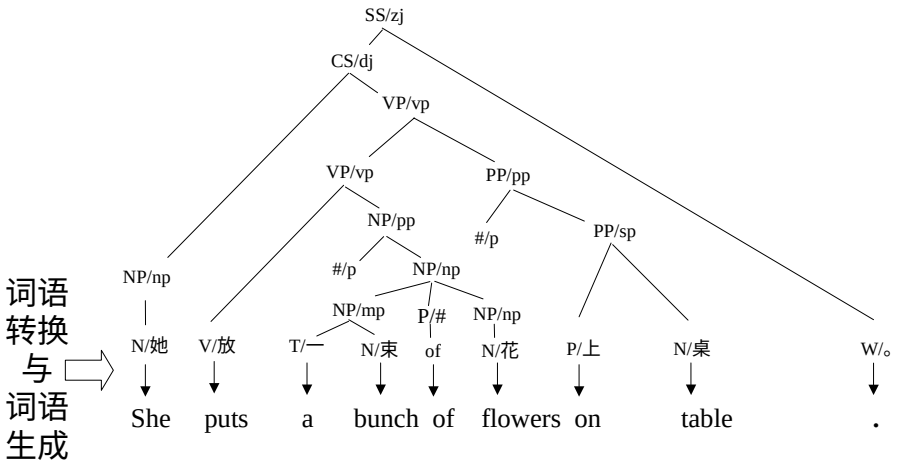
句法层面的转换方法 3



句法层面的转换方法 4



句法层面的转换方法 5



词语转换与词语生成

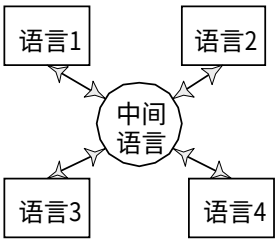
中间语言方法 1

- 利用一种中间语言（interlingua）作为翻译的中介表示形式；
- 整个翻译的过程分为“分析”和“生成”两个阶段
- 分析：源语言➔中间语言
- 生成：中间语言➔目标语言
- 分析过程只与源语言有关，与目标语言无关
- 生成过程只与目标语言有关，与源语言无关

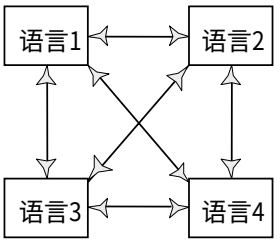
中间语言方法 2

- 中间语言方法的优点在于进行多语种翻译的时候，只需要对每种语言分别开发一个分析模块和一个生成模块，模块总数为2*n，相比之下，如果采用转换方法就需要对每两种语言之间都开发一个转换模块，模块总数为n*(n-1)

中间语言方法 3



中间语言方法



转换方法

中间语言方法 4

- 中间语言的类型
  - 自然语言：如英语、汉语
  - 人工语言：如世界语
  - 某种知识表示形式：如语义网络
- 以某种知识表示形式作为中间语言的机器翻译方法有时也称为基于知识的机器翻译方法

中间语言方法 5

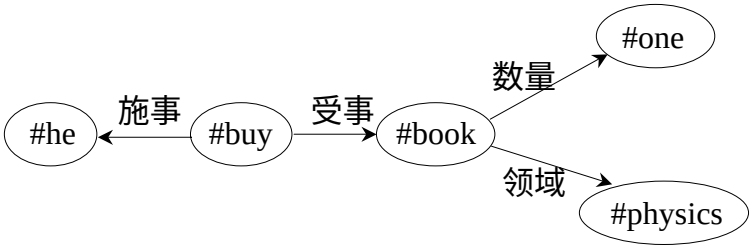
- Makoto Nagao (Kyoto University) said: ".. when the pivot language [i.e. interlingua] is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility." (Machine Translation, Oxford, 1989)
- Patel-Schneider (METAL system) said: "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." (A four-valued semantics for terminological reasoning, Artificial Intelligence, 38, 1989)

中间语言方法 6

- 基于中间语言方法一般都用于多语言的机器翻译系统中；
- 从实践看，基于中间语言的机器翻译系统还没有比较成功的先例，如日本主持的亚洲五国语言机器翻译系统，总体上是失败的；
- 目前在CSTAR多国语语音机器翻译系统中，仍然采用中间语言方法，其中间语言是一种语义表示形式，由于语音翻译都限制在非常狭窄的领域中（如机票预定），语义描述可以做到非常精确，因此采用中间语言方法有一定的合理性。

中间语言示例－语义网络

英语：He bought a book on physics.
汉语：他买了一本关于物理学的书。



说明：这里#后面表示的是概念，而不是英语词。

机器翻译的基本策略

- 基于规则的机器翻译方法
- 基于语料库的机器翻译方法
  - 基于实例的机器翻译方法
    - 基于翻译记忆的机器翻译方法
    - 基于模板（模式）的机器翻译方法
  - 基于统计的机器翻译方法
- 多引擎机器翻译方法

基于规则的方法 1

- 采用规则作为知识表示形式
  - 重叠词规则
  - 切分规则
  - 标注规则
  - 句法分析规则
  - 语义分析规则
  - 结构转换规则（产生译文句法语义结构）
  - 词语转换规则（译词选择）
  - 结构生成规则（译文结构调整）
  - 词语生成规则（译文词形生成）

基于规则的方法 2

- 优点
  - 直观，能够直接表达语言学家的知识
  - 规则的颗粒度具有很大的可伸缩性
    - 大颗粒度的规则具有很强的概括能力
    - 小颗粒度的规则具有精细的描述能力
  - 便于处理复杂的结构和进行深层次的理解，如解决长距离依赖问题
  - 系统适应性强，不依赖于具体的训练语料

基于规则的方法 3

- 缺点
  - 规则主观因素重，有时与客观事实有一定差距
  - 规则的覆盖性差，特别是细颗粒度的规则很难总结得比较全面
  - 规则之间的冲突没有好的解决办法（翘翘板现象）
  - 规则一般只局限于某一个具体的系统，规则库开发成本太高
  - 规则库的调试极其枯燥乏味

基于规则的方法－译词选择

$$ 开
  **{v} v $=[…]
    || $.主体=是,$.主体.语义类=植物
    → V<bloom> $=[…]
    || $.客体=是,$.客体.汉字=灯|机|器
    → V(!V<turn> D<on>) $=[…]
    || $.客体=是,$.客体.语义类=交通工具
    => V<drive> $=[…]
    || OTHERWISE
    => V<open> $=[…]

基于规则的方法－结构转换

&& {mp7} mp->r !mp :: $.内部结构=组合定中,…
  || %mp.定语.内部结构=单词, %mp.定语.yx=一,%mp.量词子类=集体|种类|容量|时量|度量|成形
    => NP(T/r !NP/mp) %T.TNNUM=%NP.NNUM /*这一年*/
  || %mp.定语.内部结构=单词, ,%mp.定语.yx=一,%mp.量词子类=个体
    => T(T/r M<one>) /*这一个 哪一个*/
  || %r.yx=这|那, IF %mp.定语.内部结构=单词,%mp.定语.yx=一 FALSE
    => NP(T/r !M/mp) %T.TNNUM=PLUR,$.NNUM=PLUR /*这两张*/
    => NP(T/r !NP/mp) %T.TNNUM=PLUR,$.NNUM=PLUR
  || %r.yx=~这~那,IF %mp.定语.内部结构=单词,%mp.定语.yx=一 FALSE
    => NP(T/r !M/mp) $.NNUM=%M.NNUM
    => NP(T/r !NP/mp) %T.TNSUB=%NP.NSUBC,…

基于规则的方法－结构生成

```
## { NPMP1 } NP(T !NP(T !N))
        => NP(T/T !NP/NP(!N/N))
        /* this a kind => this kind */
## { NPATN1 } NP(AP(!A) !NP(T !N))
        => P(T/T !NP/NP(AP/AP(!A/A) !N/N))
        /* red this book => this red book */
```


基于语料库的机器翻译方法

- 优点
  - 使用语料库作为翻译知识来源，无需人工编写规则，系统开发成本低，速度快
  - 从语料库中学习到的知识比较客观
  - 从语料库中学习到的知识覆盖性比较好
- 缺点
  - 系统性能依赖于语料库
  - 数据稀疏问题严重
  - 语料库中不容易活动大颗粒度的高概括性知识

基于实例的机器翻译（1）

- 长尾真(Makoto Nagao)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。
- 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。

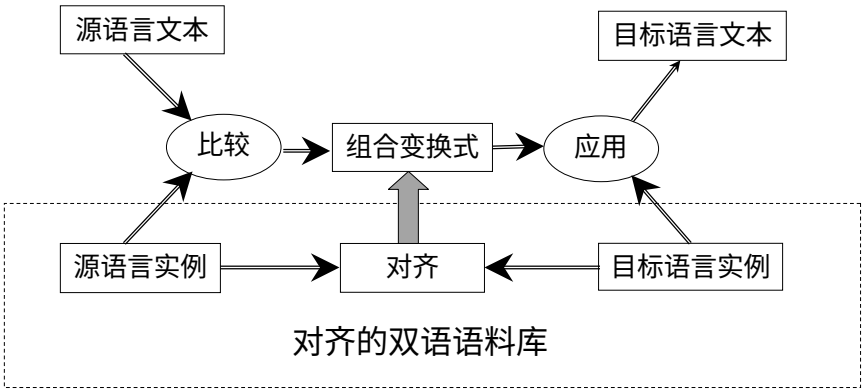
基于实例的机器翻译（2）

- 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。在翻译策略上是很有吸引力的。

基于实例的机器翻译（3）

- 优点
  - 直接使用对齐的语料库作为知识表示形式，知识库的扩充非常简单
  - 不需要进行深层次的语言分析，也可以产生高质量的译文
- 缺点
  - 覆盖率低，实用的系统需要的语料库规模极大（百万句对以上）

基于实例的机器翻译系统结构



基于实例的机器翻译－举例

要翻译句子：

(E1) He bought a book on physics.

在语料库中查到相似英语句子及其汉语译文是：

(E2) He wrote a book on history.

(C2) 他写了一本关于历史的书。

比较(E1)和(E2)两个句子，我们得到变换式：

(T1) replace(wrote, bought) and replace(history, physics)

将这个变换式中的单词都换成汉语就变成：

(T2) replace(写,买) and replace(历史,物理)

将(T2)作用于(C2)

(C1)他买了一本关于物理学的书。

基于实例的机器翻译
需要研究的问题

- 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。
- 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。
- 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。
- 开展浅层句法分析(shallow parsing)的研究：浅层句法分析以建立语段(chunk)之间的依附关系为目标，进行语段的识别，分析语段之间的依附关系。由于分析的语言单位的颗粒度比较大，歧义就比较少，有利于提高双语对齐的准确度。

基于实例的机器翻译系统

- MBT1和MBT2系统：由日本京都大学长尾真和佐藤研制。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。
- PANGLOSS系统：由美国卡内基-梅隆大学研制，这是一个多引擎机器翻译系统(Multi-engine Machine Translation)。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。
- ETOC和EBMT系统：由日本口语翻译通信研究实验室 ATR研制。 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的基于实例的日汉机器翻译系统。

复习思考题

- 访问一些知名的网上翻译网站，直观了解机器翻译
  - SYSTRAN Homepage
  - WordLingo
  - 看世界
- 尝试写一些规则，将英语句子“He wrote a book on history.”翻译成汉语句子“他写了一本关于历史的书。”
- 写一个程序实现英语数字、汉语数字和阿拉伯数字之间的互译
- 写一个程序实现英语和汉语之间时间表达式的互译